Elliptic Modular Forms and Their Applications

Don Zagier

Max-Planck-Institut für Mathematik, Vivatsgasse 7, 53111 Bonn, Germany
E-mail: zagier@mpim-bonn.mpg.de

Foreword

These notes give a brief introduction to a number of topics in the classical theory of modular forms. Some of theses topics are (planned) to be treated in much more detail in a book, currently in preparation, based on various courses held at the Collège de France in the years 2000–2004. Here each topic is treated with the minimum of detail needed to convey the main idea, and longer proofs are omitted.

Classical (or “elliptic”) modular forms are functions in the complex upper half-plane which transform in a certain way under the action of a discrete subgroup Γ of $\mathrm{SL}(2,\mathbb{R})$ such as $\mathrm{SL}(2,\mathbb{Z})$. From the point of view taken here, there are two cardinal points about them which explain why we are interested. First of all, the space of modular forms of a given weight on Γ is finite dimensional and algorithmically computable, so that it is a mechanical procedure to prove any given identity among modular forms. Secondly, modular forms occur naturally in connection with problems arising in many other areas of mathematics. Together, these two facts imply that modular forms have a huge number of applications in other fields. The principal aim of these notes – as also of the notes on Hilbert modular forms by Bruinier and on Siegel modular forms by van der Geer – is to give a feel for some of these applications, rather than emphasizing only the theory. For this reason, we have tried to give as many and as varied examples of interesting applications as possible. These applications are placed in separate mini-subsections following the relevant sections of the main text, and identified both in the text and in the table of contents by the symbol ♠. (The end of such a mini-subsection is correspondingly indicated by the symbol ♡ : these are *major* applications.) The subjects they cover range from questions of pure number theory and combinatorics to differential equations, geometry, and mathematical physics.

The notes are organized as follows. Section 1 gives a basic introduction to the theory of modular forms, concentrating on the full modular group

$\Gamma_1 = \mathrm{SL}(2,\mathbb{Z})$. Much of what is presented there can be found in standard textbooks and will be familiar to most readers, but we wanted to make the exposition self-contained. The next two sections describe two of the most important constructions of modular forms, Eisenstein series and theta series. Here too most of the material is quite standard, but we also include a number of concrete examples and applications which may be less well known. Section 4 gives a brief account of Hecke theory and of the modular forms arising from algebraic number theory or algebraic geometry whose L-series have Euler products. In the last two sections we turn to topics which, although also classical, are somewhat more specialized; here there is less emphasis on proofs and more on applications. Section 5 treats the aspects of the theory connected with differentiation of modular forms, and in particular the differential equations which these functions satisfy. This is perhaps the most important single source of applications of the theory of modular forms, ranging from irrationality and transcendence proofs to the power series arising in mirror symmetry. Section 6 treats the theory of complex multiplication. This too is a classical theory, going back to the turn of the (previous) century, but we try to emphasize aspects that are more recent and less familiar: formulas for the norms and traces of the values of modular functions at CM points, Borcherds products, and explicit Taylor expansions of modular forms. (The last topic is particularly pretty and has applications to quite varied problems of number theory.) A planned seventh section would have treated the integrals, or "periods," of modular forms, which have a rich combinatorial structure and many applications, but had to be abandoned for reasons of space and time. Apart from the first two, the sections are largely independent of one another and can be read in any order. The text contains 29 numbered "Propositions" whose proofs are given or sketched and 20 unnumbered "Theorems" which are results quoted from the literature whose proofs are too difficult (in many cases, *much* too difficult) to be given here, though in many cases we have tried to indicate what the main ingredients are. To avoid breaking the flow of the exposition, references and suggestions for further reading have not been given within the main text but collected into a single section at the end. Notations are standard (e.g., $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ for the integers, rationals, reals and complex numbers, respectively, and $\mathbb{N}$ for the strictly positive integers). Multiplication precedes division hierarchically, so that, for instance, $1/4\pi$ means $1/(4\pi)$ and not $(1/4)\pi$.

The presentation in Sections 1–5 is based partly on notes taken by Christian Grundh, Magnus Dehli Vigeland and my wife, Silke Wimmer-Zagier, of the lectures which I gave at Nordfjordeid, while that of Section 6 is partly based on the notes taken by John Voight of an earlier course on complex multiplication which I gave in Berekeley in 1992. I would like to thank all of them here, but especially Silke, who read each section of the notes as it was written and made innumerable useful suggestions concerning the exposition. And of course special thanks to Kristian Ranestad for the wonderful week in Nordfjordeid which he organized.

1 Basic Definitions

In this section we introduce the basic objects of study – the group $\mathrm{SL}(2,\mathbb{R})$ and its action on the upper half plane, the modular group, and holomorphic modular forms – and show that the space of modular forms of any weight and level is finite-dimensional. This is the key to the later applications.

1.1 Modular Groups, Modular Functions and Modular Forms

The *upper half plane*, denoted $\mathfrak{H}$, is the set of all complex numbers with positive imaginary part:

$$\mathfrak{H} \;=\; \bigl\{z \in \mathbb{C} \mid \Im(z) > 0\bigr\}\,.$$

The special linear group $\mathrm{SL}(2,\mathbb{R})$ acts on $\mathfrak{H}$ in the standard way by *Möbius transformations* (or *fractional linear transformations*):

$$\gamma \;=\; \begin{pmatrix} a & b \\ c & d \end{pmatrix} \,:\, \mathfrak{H} \to \mathfrak{H}\,, \qquad z \mapsto \gamma z \;=\; \gamma(z) \;=\; \frac{az+b}{cz+d}\,.$$

To see that this action is well-defined, we note that the denominator is non-zero and that $\mathfrak{H}$ is mapped to $\mathfrak{H}$ because, as a sinple calculation shows,

$$\Im(\gamma z) \;=\; \frac{\Im(z)}{|cz+d|^2}\,. \tag{1}$$

The transitivity of the action also follows by direct calculations, or alternatively we can view $\mathfrak{H}$ as the set of classes of $\bigl\{\left(\begin{smallmatrix}\omega_1\\ \omega_2\end{smallmatrix}\right) \in \mathbb{C}^2 \mid \omega_2 \neq 0, \Im(\omega_1/\omega_2){>}0\bigr\}$ under the equivalence relation of multiplication by a non-zero scalar, in which case the action is given by ordinary matrix multiplication from the left. Notice that the matrices $\pm\gamma$ act in the same way on $\mathfrak{H}$, so we can, and often will, work instead with the group $\mathrm{PSL}(2,\mathbb{R}) = \mathrm{SL}(2,\mathbb{R})/\{\pm 1\}$.

Elliptic modular functions and modular forms are functions in $\mathfrak{H}$ which are either invariant or transform in a specific way under the action of a discrete subgroup $\mathit{\Gamma}$ of $\mathrm{SL}(2,\mathbb{R})$. In these introductory notes we will consider only the group $\mathit{\Gamma}_1 = \mathrm{SL}(2,\mathbb{Z})$ (the “full modular group”) and its congruence subgroups (subgroups of finite index of $\mathit{\Gamma}_1$ which are described by congruence conditions on the entries of the matrices). We should mention, however, that there are other interesting discrete subgroups of $\mathrm{SL}(2,\mathbb{R})$, most notably the non-congruence subgroups of $\mathrm{SL}(2,\mathbb{Z})$, whose corresponding modular forms have rather different arithmetic properties from those on the congruence subgroups, and subgroups related to quaternion algebras over $\mathbb{Q}$, which have a compact fundamental domain. The latter are important in the study of both Hilbert and Siegel modular forms, treated in the other contributions in this volume.

The modular group takes its name from the fact that the points of the quotient space $\Gamma_1\backslash\mathfrak{H}$ are *moduli* (= parameters) for the isomorphism classes of elliptic curves over $\mathbb{C}$. To each point $z \in \mathfrak{H}$ one can associate the lattice $\Lambda_z = \mathbb{Z}.z + \mathbb{Z}.1 \subset \mathbb{C}$ and the quotient space $E_z = \mathbb{C}/\Lambda_z$, which is an elliptic curve, i.e., it is at the same time a complex curve and an abelian group. Conversely, every elliptic curve over $\mathbb{C}$ can be obtained in this way, but not uniquely: if E is such a curve, then E can be written as the quotient $\mathbb{C}/\Lambda$ for some lattice (discrete rank 2 subgroup) $\Lambda \subset \mathbb{C}$ which is unique up to "homotheties" $\Lambda \mapsto \lambda\Lambda$ with $\lambda \in \mathbb{C}^*$, and if we choose an oriented basis (ω_1, ω_2) of Λ (one with $\Im(\omega_1/\omega_2) > 0$) and use $\lambda = \omega_2^{-1}$ for the homothety, then we see that $E \cong E_z$ for some $z \in \mathfrak{H}$, but choosing a different oriented basis replaces z by γz for some $\gamma \in \Gamma_1$. The quotient space $\Gamma_1\backslash\mathfrak{H}$ is the simplest example of what is called a *moduli space*, i.e., an algebraic variety whose points classify isomorphism classes of other algebraic varieties of some fixed type. A complex-valued function on this space is called a *modular function* and, by virtue of the above discussion, can be seen as any one of four equivalent objects: a function from $\Gamma_1\backslash\mathfrak{H}$ to $\mathbb{C}$, a function $f : \mathfrak{H} \to \mathbb{C}$ satisfying the transformation equation $f(\gamma z) = f(z)$ for every $z \in \mathfrak{H}$ and every $\gamma \in \Gamma_1$, a function assigning to every elliptic curve E over $\mathbb{C}$ a complex number depending only on the isomorphism type of E, or a function on lattices in $\mathbb{C}$ satisfying $F(\lambda\Lambda) = F(\Lambda)$ for all lattices Λ and all $\lambda \in \mathbb{C}^\times$, the equivalence between f and F being given in one direction by $f(z) = F(\Lambda_z)$ and in the other by $F(\Lambda) = f(\omega_1/\omega_2)$ where (ω_1, ω_2) is any oriented basis of Λ. Generally the term "modular function", on Γ_1 or some other discrete subgroup $\Gamma \subset \mathrm{SL}(2,\mathbb{R})$, is used only for meromorphic modular functions, i.e., Γ-invariant meromorphic functions in $\mathfrak{H}$ which are of exponential growth at infinity (i.e., $f(x+iy) = \mathrm{O}(e^{Cy})$ as $y \to \infty$ and $f(x+iy) = \mathrm{O}(e^{C/y})$ as $y \to 0$ for some $C > 0$), this latter condition being equivalent to the requirement that f extends to a meromorphic function on the compactified space $\overline{\Gamma\backslash\mathfrak{H}}$ obtained by adding finitely many "cusps" to $\Gamma\backslash\mathfrak{H}$ (see below).

It turns out, however, that for the purposes of doing interesting arithmetic the modular functions are not enough and that one needs a more general class of functions called *modular forms*. The reason is that modular functions have to be allowed to be meromorphic, because there are no global holomorphic functions on a compact Riemann surface, whereas modular forms, which have a more flexible transformation behavior, are holomorphic functions (on $\mathfrak{H}$ and, in a suitable sense, also at the cusps). Every modular function can be represented as a quotient of two modular forms, and one can think of the modular functions and modular forms as in some sense the analogues of rational numbers and integers, respectively. From the point of view of functions on lattices, modular forms are simply functions $\Lambda \mapsto F(\Lambda)$ which transform under homotheties by $F(\lambda\Lambda) = \lambda^{-k}F(\Lambda)$ rather than simply by $F(\lambda\Lambda) = F(\Lambda)$ as before, where k is a fixed integer called the *weight* of the modular form. If we translate this back into the language of functions on $\mathfrak{H}$ via $f(z) = F(\Lambda_z)$ as before, then we see that f is now required to satisfy the *modular transformation property*

$$f\left(\frac{az+b}{cz+d}\right) \;=\; (cz+d)^k\, f(z) \tag{2}$$

for all $z \in \mathfrak{H}$ and all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \varGamma_1$; conversely, given a function $f : \mathfrak{H} \to \mathbb{C}$ satisfying (2), we can define a funcion on lattices, homogeneous of degree $-k$ with respect to homotheties, by $F(\mathbb{Z}.\omega_1 + \mathbb{Z}.\omega_2) = \omega_2^{-k} f(\omega_1/\omega_2)$. As with modular functions, there is a standard convention: when the word "modular form" (on some discrete subgroup $\varGamma$ of $\mathrm{SL}(2,\mathbb{R})$) is used with no further adjectives, one generally means "holomorphic modular form", i.e., a function f on $\mathfrak{H}$ satisfying (2) for all $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \varGamma$ which is holomorphic in $\mathfrak{H}$ and of subexponential growth at infinity (i.e., f satisfies the same estimate as above, but now for *all* rather than *some* $C > 0$). This growth condition, which corresponds to holomorphy at the cusps in a sense which we do not explain now, implies that the growth at infinity is in fact polynomial; more precisely, f automatically satisfies $f(z) = \mathrm{O}(1)$ as $y \to \infty$ and $f(x+iy) = \mathrm{O}(y^{-k})$ as $y \to 0$. We denote by $M_k(\varGamma)$ the space of holomorphic modular forms of weight k on $\varGamma$. As we will see in detail for $\varGamma = \varGamma_1$, this space is finite-dimensional, effectively computable for all k, and zero for $k < 0$, and the algebra $M_*(\varGamma) := \bigoplus_k M_k(\varGamma)$ of all modular forms on $\varGamma$ is finitely generated over $\mathbb{C}$.

If we specialize (2) to the matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$, which belongs to $\varGamma_1$, then we see that any modular form on $\varGamma_1$ satisfies $f(z+1) = f(z)$ for all $z \in \mathfrak{H}$, i.e., it is a periodic function of period 1. It is therefore a function of the quantity $e^{2\pi i z}$, traditionally denoted q; more precisely, we have the *Fourier development*

$$f(z) \;=\; \sum_{n=0}^{\infty} a_n\, e^{2\pi i n z} \;=\; \sum_{n=0}^{\infty} a_n\, q^n \qquad \bigl(z \in \mathfrak{H},\;\; q \;=\; e^{2\pi i z}\bigr)\,, \tag{3}$$

where the fact that only terms q^n with $n \ge 0$ occur is a consequence of (and in the case of $\varGamma_1$, in fact equivalent to) the growth conditions on f just given. It is this Fourier development which is responsible for the great importance of modular forms, because it turns out that there are many examples of modular forms f for which the Fourier coefficients a_n in (3) are numbers that are of interest in other domains of mathematics.

1.2 The Fundamental Domain of the Full Modular Group

In the rest of §1 we look in more detail at the modular group. Because $\varGamma_1$ contains the element $-1 = \left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$ which fixes every point of $\mathfrak{H}$, we can also consider the action of the quotient group $\overline{\varGamma}_1 = \varGamma_1/\{\pm 1\} = \mathrm{PSL}(2,\mathbb{Z}) \subset \mathrm{PSL}(2,\mathbb{R})$ on $\mathfrak{H}$. It is clear from (2) that a modular form of odd weight on $\varGamma_1$ (or on any subgroup of $\mathrm{SL}(2,\mathbb{R})$ containing -1) must vanish, so we can restrict our attention to even k. But then the "automorphy factor" $(cz+d)^k$ in (2) is unchanged when we replace $\gamma \in \varGamma_1$ by $-\gamma$, so that we can consider equation (2) for k even and $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \overline{\varGamma}_1$. By a slight abuse of notation, we will use the same notation for an element γ of $\varGamma_1$ and its image $\pm\gamma$ in $\overline{\varGamma}_1$, and, for k even, will not distinguish between the isomorphic spaces $M_k(\varGamma_1)$ and $M_k(\overline{\varGamma}_1)$.

The group $\overline{\Gamma}_1$ is generated by the two elements $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $S = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, with the relations $S^2 = (ST)^3 = 1$. The actions of S and T on $\mathfrak{H}$ are given by

$$S\,:\,z \mapsto -1/z\,, \qquad T\,:\,z \mapsto z+1\,.$$

Therefore f is a modular form of weight k on Γ_1 precisely when f is periodic with period 1 and satisfies the single further functional equation

$$f\bigl(-1/z\bigr) \;=\; z^k f(z) \qquad (z \in \mathfrak{H})\,. \tag{4}$$

If we know the value of a modular form f on some group Γ at one point $z \in \mathfrak{H}$, then equation (2) tells us the value at all points in the same Γ_1-orbit as z. So to be able to completely determine f it is enough to know the value at one point from each orbit. This leads to the concept of a *fundamental domain* for Γ, namely an open subset $\mathcal{F} \subset \mathfrak{H}$ such that no two distinct points of $\mathcal{F}$ are equivalent under the action of Γ and every point $z \in \mathfrak{H}$ is Γ-equivalent to some point in the closure $\overline{\mathcal{F}}$ of $\mathcal{F}$.

Proposition 1. *The set*

$$\mathcal{F}_1 \;=\; \bigl\{z \in \mathfrak{H} \;\mid\; |z| > 1,\; |\Re(z)| \,<\, \tfrac{1}{2}\bigr\}$$

is a fundamental domain for the full modular group Γ_1. (See Fig. 1A.)

Proof. Take a point $z \in \mathfrak{H}$. Then $\{mz+n \mid m, n \in \mathbb{Z}\}$ is a lattice in $\mathbb{C}$. Every lattice has a point different from the origin of minimal modulus. Let $cz+d$ be such a point. The integers c, d must be relatively prime (otherwise we could divide $cz+d$ by an integer to get a new point in the lattice of even smaller modulus). So there are integers a and b such that $\gamma_1 = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_1$. By the transformation property (1) for the imaginary part $y = \Im(z)$ we get that $\Im(\gamma_1 z)$ is a maximal member of $\{\Im(\gamma z) \mid \gamma \in \Gamma_1\}$. Set $z^* = T^n\gamma_1 z = \gamma_1 z + n$,

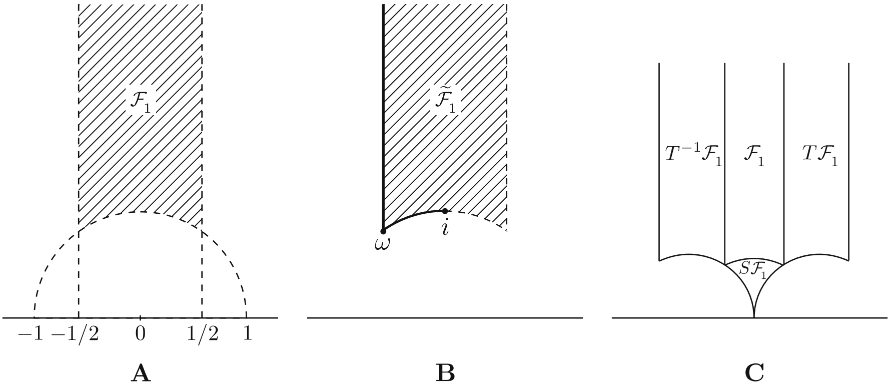

Fig. 1. The standard fundamental domain for $\overline{\Gamma}_1$ and its neighbors

where n is such that $|\Re(z^*)| \le \frac{1}{2}$. We cannot have $|z^*| < 1$, because then we would have $\Im(-1/z^*) = \Im(z^*)/|z^*|^2 > \Im(z^*)$ by (1), contradicting the maximality of $\Im(z^*)$. So $z^* \in \overline{\mathcal{F}_1}$, and z is equivalent under $\mathit{\Gamma}_1$ to z^*.

Now suppose that we had two $\mathit{\Gamma}_1$-equivalent points z_1 and $z_2 = \gamma z_1$ in $\mathcal{F}_1$, with $\gamma \ne \pm 1$. This γ cannot be of the form T^n since this would contradict the condition $|\Re(z_1)|$, $|\Re(z_2)| < \frac{1}{2}$, so $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with $c \ne 0$. Note that $\Im(z) > \sqrt{3}/2$ for all $z \in \mathcal{F}_1$. Hence from (1) we get

$$\frac{\sqrt{3}}{2} \;<\; \Im(z_2) \;=\; \frac{\Im(z_1)}{|cz_1 + d|^2} \le \frac{\Im(z_1)}{c^2\,\Im(z_1)^2} < \frac{2}{c^2\sqrt{3}}\,,$$

which can only be satisfied if $c = \pm 1$. Without loss of generality we may assume that $\Im z_1 \le \Im z_2$. But $|\pm z_1 + d| \ge |z_1| > 1$, and this gives a contradiction with the transformation property (1).

Remarks. 1. The points on the borders of the fundamental region are $\mathit{\Gamma}_1$-equivalent as follows: First, the points on the two lines $\Re(z) = \pm\frac{1}{2}$ are equivalent by the action of $T : z \mapsto z + 1$. Secondly, the points on the left and right halves of the arc $|z| = 1$ are equivalent under the action of $S : z \mapsto -1/z$. In fact, these are the only equivalences for the points on the boundary. For this reason we define $\widetilde{\mathcal{F}_1}$ to be the semi-closure of $\mathcal{F}_1$ where we have added only the boundary points with non-positive real part (see Fig. 1B). Then every point of $\mathfrak{H}$ is $\mathit{\Gamma}_1$-equivalent to a *unique* point of $\widetilde{\mathcal{F}_1}$, i.e., $\widetilde{\mathcal{F}_1}$ is a *strict fundamental domain* for the action of $\mathit{\Gamma}_1$. (But terminology varies, and many people use the words "fundamental domain" for the strict fundamental domain or for its closure, rather than for the interior.)

2. The description of the fundamental domain $\mathcal{F}_1$ also implies the above-mentioned fact that $\mathit{\Gamma}_1$ (or $\overline{\mathit{\Gamma}}_1$) is generated by S and T. Indeed, by the very definition of a fundamental domain we know that $\overline{\mathcal{F}_1}$ and its translates $\gamma\overline{\mathcal{F}_1}$ by elements γ of $\mathit{\Gamma}_1$ cover $\mathfrak{H}$, disjointly except for their overlapping boundaries (a so-called "tesselation" of the upper half-plane). The neighbors of $\mathcal{F}_1$ are $T^{-1}\mathcal{F}_1$, $S\mathcal{F}_1$ and $T\mathcal{F}_1$ (see Fig. 1C), so one passes from any translate $\gamma\mathcal{F}_1$ of $\mathcal{F}_1$ to one of its three neighbors by applying $\gamma S \gamma^{-1}$ or $\gamma T^{\pm 1}\gamma^{-1}$. In particular, if the element γ describing the passage from $\mathcal{F}_1$ to a given translated fundamental domain $\mathcal{F}_1' = \gamma\mathcal{F}_1$ can be written as a word in S and T, then so can the element of $\mathit{\Gamma}_1$ which describes the motion from $\mathcal{F}_1$ to any of the neighbors of $\mathcal{F}_1'$. Therefore by moving from neighbor to neighbor across the whole upper half-plane we see inductively that this property holds for every $\gamma \in \mathit{\Gamma}_1$, as asserted. More generally, one sees that if one has given a fundamental domain $\mathcal{F}$ for any discrete group $\mathit{\Gamma}$, then the elements of $\mathit{\Gamma}$ which identify in pairs the sides of $\overline{\mathcal{F}}$ always generate $\mathit{\Gamma}$.

♠ Finiteness of Class Numbers

Let D be a negative discriminant, i.e., a negative integer which is congruent to 0 or 1 modulo 4. We consider binary quadratic forms of the form $Q(x, y) =$

$Ax^2 + Bxy + Cy^2$ with A, B, $C \in \mathbb{Z}$ and $B^2 - 4AC = D$. Such a form is definite (i.e., $Q(x, y) \neq 0$ for non-zero $(x, y) \in \mathbb{R}^2$) and hence has a fixed sign, which we take to be positive. (This is equivalent to $A > 0$.) We also assume that Q is primitive, i.e., that $\gcd(A, B, C) = 1$. Denote by $\mathfrak{Q}_D$ the set of these forms. The group Γ_1 (or indeed $\overline{\Gamma}_1$) acts on $\mathfrak{Q}_D$ by $Q \mapsto Q \circ \gamma$, where $(Q \circ \gamma)(x, y) = Q(ax + by, cx + dy)$ for $\gamma = \pm\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \overline{\Gamma_1}$. We claim that the number of equivalence classes under this action is finite. This number, called the *class number* of D and denoted $h(D)$, also has an interpretation as the number of ideal classes (either for the ring of integers or, if D is a non-trivial square multiple of some other discriminant, for a non-maximal order) in the imaginary quadratic field $\mathbb{Q}(\sqrt{D})$, so this claim is a special case – historically the first one, treated in detail in Gauss's *Disquisitiones Arithmeticae* – of the general theorem that the number of ideal classes in any number field is finite. To prove it, we observe that we can associate to any $Q \in \mathfrak{Q}_D$ the unique root $\mathfrak{z}_Q = (-B + \sqrt{D})/2A$ of $Q(\mathfrak{z}, 1) = 0$ in the upper half-plane (here $\sqrt{D} = +i\sqrt{|D|}$ by definition and $A > 0$ by assumption). One checks easily that $\mathfrak{z}_{Q \circ \gamma} = \gamma^{-1}(\mathfrak{z}_Q)$ for any $\gamma \in \overline{\Gamma}_1$, so each $\overline{\Gamma}_1$-equivalence class of forms $Q \in \mathfrak{Q}_D$ has a unique representative belonging to the set

$$\mathfrak{Q}_D^{\text{red}} = \left\{ [A, B, C] \in \mathfrak{Q}_D \mid -A < B \leq A < C \quad \text{or} \quad 0 \leq B \leq A = C \right\} \tag{5}$$

of $Q \in \mathfrak{Q}_D$ for which $\mathfrak{z}_Q \in \widetilde{\mathcal{F}_1}$ (the so-called *reduced* quadratic forms of discriminant D), and this set is finite because $C \geq A \geq |B|$ implies $|D| = 4AC - B^2 \geq 3A^2$, so that both A and B are bounded in absolute value by $\sqrt{|D|/3}$, after which C is fixed by $C = (B^2 - D)/4A$. This even gives us a way to compute $h(D)$ effectively, e.g., $\mathfrak{Q}_{-47}^{\text{red}} = \{[1, 1, 12],\ [2, \pm 1, 6],\ [3, \pm 1, 4]\}$ and hence $h(-47) = 5$. We remark that the class numbers $h(D)$, or a small modification of them, are themselves the coefficients of a modular form (of weight 3/2), but this will not be discussed further in these notes. ♡

1.3 The Finite Dimensionality of $M_k(\Gamma)$

We end this section by applying the description of the fundamental domain to show that $M_k(\Gamma_1)$ is finite-dimensional for every k and to get an upper bound for its dimension. In §2 we will see that this upper bound is in fact the correct value.

If f is a modular form of weight k on Γ_1 or any other discrete group Γ, then f is not a well-defined function on the quotient space $\Gamma \backslash \mathfrak{H}$, but the transformation formula (2) implies that the order of vanishing $\text{ord}_z(f)$ at a point $z \in \mathfrak{H}$ depends only on the orbit Γz. We can therefore define a local order of vanishing, $\text{ord}_P(f)$, for each $P \in \Gamma \backslash \mathfrak{H}$. The key assertion is that the total number of zeros of f, i.e., the sum of all of these local orders, depends only on Γ and k. But to make this true, we have to look more carefully at the geometry of the quotient space $\Gamma \backslash \mathfrak{H}$, taking into account the fact that some points (the so-called *elliptic fixed points*, corresponding to the points

$z \in \mathfrak{H}$ which have a non-trivial stabilizer for the image of Γ in $\mathrm{PSL}(2,\mathbb{R})$) are singular and also that $\Gamma\backslash\mathfrak{H}$ is not compact, but has to be compactified by the addition of one or more further points called *cusps*. We explain this for the case $\Gamma = \Gamma_1$.

In §1.2 we identified the quotient space $\Gamma_1\backslash\mathfrak{H}$ as a set with the semi-closure $\widetilde{\mathcal{F}_1}$ of $\mathcal{F}_1$ and as a topological space with the quotient of $\overline{\mathcal{F}_1}$ obtained by identifying the opposite sides (lines $\Re(z) = \pm\frac{1}{2}$ or halves of the arc $|z| = 1$) of the boundary $\partial\overline{\mathcal{F}_1}$. For a generic point of $\widetilde{\mathcal{F}_1}$ the stabilizer subgroup of $\overline{\Gamma}_1$ is trivial. But the two points $\omega = \frac{1}{2}(-1 + i\sqrt{3}) = e^{2\pi i/3}$ and i are stabilized by the cyclic subgroups of order 3 and 2 generated by ST and S respectively. This means that in the quotient manifold $\Gamma_1\backslash\mathfrak{H}$, ω and i are singular. (From a metric point of view, they have neighborhoods which are not discs, but quotients of a disc by these cyclic subgroups, with total angle 120° or 180° instead of 360°.) If we define an integer n_P for every $P \in \Gamma_1\backslash\mathfrak{H}$ as the order of the stabilizer in $\overline{\Gamma}_1$ of any point in $\mathfrak{H}$ representing P, then n_P equals 2 or 3 if P is Γ_1-equivalent to i or ω and $n_P = 1$ otherwise. We also have to consider the compactified quotient $\overline{\Gamma_1\backslash\mathfrak{H}}$ obtained by adding a point at infinity ("cusp") to $\Gamma_1\backslash\mathfrak{H}$. More precisely, for $Y > 1$ the image in $\Gamma_1\backslash\mathfrak{H}$ of the part of $\mathfrak{H}$ above the line $\Im(z) = Y$ can be identified via $q = e^{2\pi i z}$ with the punctured disc $0 < q < e^{-2\pi Y}$. Equation (3) tells us that a holomorphic modular form of any weight k on Γ_1 is not only a well-defined function on this punctured disc, but extends holomorphically to the point $q = 0$. We therefore define $\overline{\Gamma_1\backslash\mathfrak{H}} = \Gamma_1\backslash\mathfrak{H} \cup \{\infty\}$, where the point "$\infty$" corresponds to $q = 0$, with q as a local parameter. One can also think of $\overline{\Gamma_1\backslash\mathfrak{H}}$ as the quotient of $\overline{\mathfrak{H}}$ by Γ_1, where $\overline{\mathfrak{H}} = \mathfrak{H} \cup \mathbb{Q} \cup \{\infty\}$ is the space obtained by adding the full Γ_1-orbit $\mathbb{Q} \cup \{\infty\}$ of ∞ to $\mathfrak{H}$. We define the *order of vanishing at infinity* of f, denoted $\mathrm{ord}_\infty(f)$, as the smallest integer n such that $a_n \neq 0$ in the Fourier expansion (3).

Proposition 2. *Let f be a non-zero modular form of weight k on Γ_1. Then*

$$\sum_{P \in \Gamma_1\backslash\mathfrak{H}} \frac{1}{n_P}\, \mathrm{ord}_P(f) \;+\; \mathrm{ord}_\infty(f) \;=\; \frac{k}{12}\,. \tag{6}$$

Proof. Let D be the closed set obtained from $\overline{\mathcal{F}_1}$ by deleting ε-neighborhoods of all zeros of f and also the "neighborhood of infinity" $\Im(z) > Y = \varepsilon^{-1}$, where ε is chosen sufficiently small that all of these neighborhoods are disjoint (see Fig. 2.) Since f has no zeros in D, Cauchy's theorem implies that the integral of $d\bigl(\log f(z)\bigr) = \dfrac{f'(z)}{f(z)}\,dz$ over the boundary of D is 0. This boundary consists of several parts: the horizontal line from $-\frac{1}{2} + iY$ to $\frac{1}{2} + iY$, the two vertical lines from ω to $-\frac{1}{2} + iY$ and from $\omega + 1$ to $\frac{1}{2} + iY$ (with some ε-neighborhoods removed), the arc of the circle $|z| = 1$ from ω to $\omega + 1$ (again with some ε-neighborhoods deleted), and the boundaries of the ε-neighborhoods of the zeros P of f. These latter have total angle 2π if P is not an elliptic fixed point

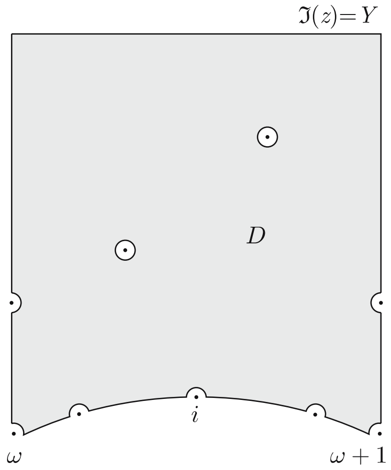

Fig. 2. The zeros of a modular form

(they consist of a full circle if P is an interior point of $\overline{\mathcal{F}_1}$ and of two half-circles if P corresponds to a boundary point of $\overline{\mathcal{F}_1}$ different from ω, $\omega+1$ or i), and total angle π or $2\pi/3$ if $P \sim i$ or ω. The corresponding contributions to the integral are as follows. The two vertical lines together give 0, because f takes on the same value on both and the orientations are opposite. The horizontal line from $-\frac{1}{2}+iY$ to $\frac{1}{2}+iY$ gives a contribution $2\pi i\,\mathrm{ord}_\infty(f)$, because $d(\log f)$ is the sum of $\mathrm{ord}_\infty(f)\,dq/q$ and a function of q which is holomorphic at 0, and this integral corresponds to an integral around a small circle $|q| = e^{-2\pi Y}$ around $q=0$. The integral on the boundary of the deleted ε-neighborhood of a zero P of f contributes $2\pi i\,\mathrm{ord}_P(f)$ if $n_P=1$ by Cauchy's theorem, because $\mathrm{ord}_P(f)$ is the residue of $d(\log f(z))$ at $z=P$, while for $n_P>1$ we must divide by n_P because we are only integrating over one-half or one-third of the full circle around P. Finally, the integral along the bottom arc contributes $\pi ik/6$, as we see by breaking up this arc into its left and right halves and applying the formula $d\log f(Sz) = d\log f(z) + kdz/z$, which is a consequence of the transformation equation (4). Combining all of these terms with the appropriate signs dictated by the orientation, we obtain (6). The details are left to the reader.

Corollary 1. *The dimension of $M_k(\Gamma_1)$ is 0 for $k<0$ or k odd, while for even $k \ge 0$ we have*

$$\dim M_k(\Gamma_1) \;\le\; \begin{cases} [k/12]+1 & \text{if } k \not\equiv 2 \pmod{12} \\ [k/12] & \text{if } k \equiv 2 \pmod{12}\,. \end{cases} \tag{7}$$

Proof. Let $m = [k/12]+1$ and choose m distinct non-elliptic points $P_i \in \Gamma_1\backslash\mathfrak{H}$. Given any modular forms $f_1,\ldots,f_{m+1} \in M_k(\Gamma_1)$, we can find a linear combination f of them which vanishes in all P_i, by linear algebra. But then $f \equiv 0$ by the proposition, since $m > k/12$, so the f_i are linearly dependent. Hence $\dim M_k(\Gamma_1) \le m$. If $k \equiv 2 \pmod{12}$ we can improve the estimate by 1 by noticing that the only way to satisfy (6) is to have (at least) a simple zero at i and a double zero at ω (contributing a total of $1/2 + 2/3 = 7/6$ to $\sum \mathrm{ord}_P(f)/n_P$) together with $k/12 - 7/6 = m - 1$ further zeros, so that the same argument now gives $\dim M_k(\Gamma_1) \le m - 1$.

Corollary 2. *The space $M_{12}(\Gamma_1)$ has dimension ≤ 2, and if f, $g \in M_{12}(\Gamma_1)$ are linearly independent, then the map $z \mapsto f(z)/g(z)$ gives an isomorphism from $\Gamma_1\backslash\mathfrak{H} \cup \{\infty\}$ to $\mathbb{P}^1(\mathbb{C})$.*

Proof. The first statement is a special case of Corollary 1. Suppose that f and g are linearly independent elements of $M_{12}(\Gamma_1)$. For any $(0,0) \ne (\lambda, \mu) \in \mathbb{C}^2$ the modular form $\lambda f - \mu g$ of weight 12 has exactly one zero in $\Gamma_1\backslash\mathfrak{H} \cup \{\infty\}$ by Proposition 2, so the modular function $\psi = f/g$ takes on every value $(\mu : \lambda) \in \mathbb{P}^1(\mathbb{C})$ exactly once, as claimed.

We will make an explicit choice of f, g and ψ in §2.4, after we have introduced the “discriminant function” $\Delta(z) \in M_{12}(\Gamma_1)$.

The true interpretation of the factor $1/12$ multiplying k in equation (6) is as $1/4\pi$ times the volume of $\Gamma_1\backslash\mathfrak{H}$, taken with respect to the hyperbolic metric. We say only a few words about this, since these ideas will not be used again. To give a metric on a manifold is to specify the distance between any two sufficiently near points. The *hyperbolic metric* in $\mathfrak{H}$ is defined by saying that the hyperbolic distance between two points in a small neighborhood of a point $z = x + iy \in \mathfrak{H}$ is very nearly $1/y$ times the Euclidean distance between them, so the volume element, which in Euclidean geometry is given by the 2-form $dx\,dy$, is given in hyperbolic geometry by $d\mu = y^{-2}dx\,dy$. Thus

$$\begin{aligned} \mathrm{Vol}\big(\Gamma_1\backslash\mathfrak{H}\big) &= \int_{\mathcal{F}_1} d\mu = \int_{-1/2}^{1/2} \left(\int_{\sqrt{1-x^2}}^{\infty} \frac{dy}{y^2} \right) dx \\ &= \int_{-1/2}^{1/2} \frac{dx}{\sqrt{1-x^2}} = \arcsin(x)\Big|_{-1/2}^{1/2} = \frac{\pi}{3}\,. \end{aligned}$$

Now we can consider other discrete subgroups of $\mathrm{SL}(2,\mathbb{R})$ which have a fundamental domain of finite volume. (Such groups are usually called *Fuchsian groups of the first kind*, and sometimes “lattices”, but we will reserve this latter term for discrete cocompact subgroups of Euclidean spaces.) Examples are the subgroups $\Gamma \subset \Gamma_1$ of finite index, for which the volume of $\Gamma\backslash\mathfrak{H}$ is $\pi/3$ times the index of Γ in Γ_1 (or more precisely, of the image of Γ in $\mathrm{PSL}(2,\mathbb{R})$ in $\overline{\Gamma}_1$). If Γ is any such group, then essentially the same proof as for Proposition 2 shows that the number of Γ-inequivalent zeros of any non-zero modular form $f \in M_k(\Gamma)$ equals $k\,\mathrm{Vol}(\Gamma\backslash\mathfrak{H})/4\pi$, where just as in the case of Γ_1 we must count the zeros at elliptic fixed points or cusps of Γ with appropriate multiplicities. The same argument as for Corollary 1 of Proposition 2 then tells us $M_k(\Gamma)$ is finite dimensional and gives an explicit upper bound:

Proposition 3. *Let Γ be a discrete subgroup of $SL(2,\mathbb{R})$ for which $\Gamma\backslash\mathfrak{H}$ has finite volume V. Then* $\dim M_k(\Gamma) \le \dfrac{kV}{4\pi} + 1$ *for all $k \in \mathbb{Z}$.*

In particular, we have $M_k(\Gamma) = \{0\}$ for $k < 0$ and $M_0(\Gamma) = \mathbb{C}$, i.e., there are no holomorphic modular forms of negative weight on any group Γ, and the only modular forms of weight 0 are the constants. A further consequence is that any three modular forms on Γ are algebraically dependent. (If f, g, h were algebraically independent modular forms of positive weights, then for large k the dimension of $M_k(\Gamma)$ would be at least the number of monomials in f, g, h of total weight k, which is bigger than some positive multiple of k^2, contradicting the dimension estimate given in the proposition.) Equivalently, any two modular functions on Γ are algebraically dependent, since every modular function is a quotient of two modular forms. This is a special case of the general fact that there cannot be more than n algebraically independent algebraic functions on an algebraic variety of dimension n. But the most important consequence of Proposition 3 from our point of view is that it is the origin of the (unreasonable?) effectiveness of modular forms in number theory: if we have two interesting arithmetic sequences $\{a_n\}_{n\ge0}$ and $\{b_n\}_{n\ge0}$ and conjecture that they are identical (and clearly many results of number theory can be formulated in this way), then if we can show that both $\sum a_n q^n$ and $\sum b_n q^n$ are modular forms of the same weight and group, we need only verify the equality $a_n = b_n$ for a finite number of n in order to know that it is true in general. There will be many applications of this principle in these notes.

2 First Examples: Eisenstein Series and the Discriminant Function

In this section we construct our first examples of modular forms: the Eisenstein series $E_k(z)$ of weight $k > 2$ and the discriminant function $\Delta(z)$ of weight 12, whose definition is closely connected to the non-modular Eisenstein series $E_2(z)$.

2.1 Eisenstein Series and the Ring Structure of $M_*(\Gamma_1)$

There are two natural ways to introduce the Eisenstein series. For the first, we observe that the characteristic transformation equation (2) of a modular form can be written in the form $f|_k\gamma = f$ for $\gamma \in \Gamma$, where $f|_k\gamma : \mathfrak{H} \to \mathbb{C}$ is defined by

$$\bigl(f\,\bigl|_k g\bigr)(z) \;=\; (cz+d)^{-k}\, f\left(\frac{az+b}{cz+d}\right) \qquad \bigl(z \in \mathbb{C}, \quad g \;=\; \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{R})\bigr)\,. \tag{8}$$

One checks easily that for fixed $k \in \mathbb{Z}$, the map $f \mapsto f|_k g$ defines an operation of the group $\mathrm{SL}(2,\mathbb{R})$ (i.e., $f|_k(g_1g_2) = (f|_kg_1)|_kg_2$ for all g_1, $g_2 \in \mathrm{SL}(2,\mathbb{R})$) on the vector space of holomorphic functions in $\mathfrak{H}$ having subexponential or polynomial growth. The space $M_k(\Gamma)$ of holomorphic modular forms of weight k on a group $\Gamma \subset \mathrm{SL}(2,\mathbb{R})$ is then simply the subspace of this vector space fixed by Γ.

If we have a linear action $v \mapsto v|g$ of a *finite* group G on a vector space V, then an obvious way to construct a G-invariant vector in V is to start with an arbitrary vector $v_0 \in V$ and form the sum $v = \sum_{g\in G} v_0|g$ (and to hope that the result is non-zero). If the vector v_0 is invariant under some subgroup $G_0 \subset G$, then the vector $v_0|g$ depends only on the coset $G_0g \in G_0\backslash G$ and we can form instead the smaller sum $v = \sum_{g\in G_0\backslash G} v_0|g$, which again is G-invariant. If G is infinite, the same method sometimes applies, but we now have to be careful about convergence. If the vector v_0 is fixed by an infinite subgroup G_0 of G, then this improves our chances because the sum over $G_0\backslash G$ is much smaller than a sum over all of G (and in any case $\sum_{g\in G} v|g$ has no chance of converging since every term occurs infinitely often). In the context when $G = \Gamma \subset \mathrm{SL}(2,\mathbb{R})$ is a Fuchsian group (acting by $|_k$) and v_0 a rational function, the modular forms obtained in this way are called *Poincaré series.* An especially easy case is that when v_0 is the constant function “1” and $\Gamma_0 = \Gamma_\infty$, the stabilizer of the cusp at infinity. In this case the series $\sum_{\Gamma_\infty\backslash\Gamma} 1|_k\gamma$ is called an *Eisenstein series.*

Let us look at this series more carefully when $\Gamma = \Gamma_1$. A matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{R})$ sends ∞ to a/c, and hence belongs to the stabilizer of ∞ if and only if $c = 0$. In Γ_1 these are the matrices $\pm\left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right)$ with $n \in \mathbb{Z}$, i.e., up to sign the matrices T^n. We can assume that k is even (since there are no modular forms of odd weight on Γ_1) and hence work with $\overline{\Gamma_1} = \mathrm{PSL}(2,\mathbb{Z})$, in which case the stabilizer $\overline{\Gamma}_\infty$ is the infinite cyclic group generated by T. If we multiply an arbitrary matrix $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ on the left by $\left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right)$, then the resulting matrix $\gamma' = \left(\begin{smallmatrix} a+nc & b+nd \\ c & d \end{smallmatrix}\right)$ has the same bottom row as γ. Conversely, if $\gamma' = \left(\begin{smallmatrix} a' & b' \\ c & d \end{smallmatrix}\right) \in \Gamma_1$ has the same bottom row as γ, then from $(a'-a)d-(b'-b)c = \det(\gamma)-\det(\gamma') = 0$ and $(c,d) = 1$ (the elements of any row or column of a matrix in $\mathrm{SL}(2,\mathbb{Z})$ are coprime!) we see that $a'-a = nc$, $b'-b = nd$ for some $n \in \mathbb{Z}$, i.e., $\gamma' = T^n\gamma$. Since every coprime pair of integers occurs as the bottom row of a matrix in $\mathrm{SL}(2,\mathbb{Z})$, these considerations give the formula

$$E_k(z) \;=\; \sum_{\gamma\in\Gamma_\infty\backslash\Gamma_1} 1\bigl|_k\gamma \;=\; \sum_{\gamma\in\overline{\Gamma}_\infty\backslash\overline{\Gamma}_1} 1\bigl|_k\gamma \;=\; \frac{1}{2}\sum_{\substack{c,\,d\in\mathbb{Z} \\ (c,d)\,=\,1}} \frac{1}{(cz+d)^k} \tag{9}$$

for the Eisenstein series (the factor $\frac{1}{2}$ arises because $(c\ d)$ and $(-c\ -d)$ give the same element of $\Gamma_1\backslash\overline{\Gamma}_1$). It is easy to see that this sum is absolutely convergent for $k > 2$ (the number of pairs (c, d) with $N \le |cz + d| < N + 1$ is the number of lattice points in an annulus of area $\pi(N + 1)^2 - \pi N^2$ and hence is $\mathrm{O}(N)$, so the series is majorized by $\sum_{N=1}^{\infty} N^{1-k}$), and this absolute convergence guarantees the modularity (and, since it is locally uniform in z, also the holomorphy) of the sum. The function $E_k(z)$ is therefore a modular form of weight k for all even $k \ge 4$. It is also clear that it is non-zero, since for $\mathfrak{I}(z) \to \infty$ all the terms in (9) except $(c\ d) = (\pm 1\ 0)$ tend to 0, the convergence of the series being sufficiently uniform that their sum also goes to 0 (left to the reader), so $E_k(z) = 1 + \mathrm{o}(1) \ne 0$.

The second natural way of introducing the Eisenstein series comes from the interpretation of modular forms given in the beginning of §1.1, where we identified solutions of the transformation equation (2) with functions on lattices $\Lambda \subset \mathbb{C}$ satisfying the homogeneity condition $F(\lambda\Lambda) = \lambda^{-k}F(\Lambda)$ under homotheties $\Lambda \mapsto \lambda\Lambda$. An obvious way to produce such a homogeneous function – if the series converges – is to form the sum $G_k(\Lambda) = \frac{1}{2}\sum_{\lambda\in\Lambda\smallsetminus 0} \lambda^{-k}$ of the $(-k)$th powers of the non-zero elements of Λ. (The factor "$\frac{1}{2}$" has again been introduce to avoid counting the vectors λ and $-\lambda$ doubly when k is even; if k is odd then the series vanishes anyway.) In terms of $z \in \mathfrak{H}$ and its associated lattice $\Lambda_z = \mathbb{Z}.z + \mathbb{Z}.1$, this becomes

$$G_k(z) \;=\; \frac{1}{2} \sum_{\substack{m,\,n\in\mathbb{Z}\\ (m,n)\ne(0,0)}} \frac{1}{(mz + n)^k} \qquad (k > 2,\ z \in \mathfrak{H})\,, \tag{10}$$

where the sum is again absolutely and locally uniformly convergent for $k > 2$, guaranteeing that $G_k \in M_k(\Gamma_1)$. The modularity can also be seen directly by noting that $(G_k|_k\gamma)(z) = \sum_{m,n}(m'z + n')^{-k}$ where $(m', n') = (m, n)\gamma$ runs over the non-zero vectors of $\mathbb{Z}^2 \smallsetminus \{(0, 0)\}$ as (m, n) does.

In fact, the two functions (9) and (10) are proportional, as is easily seen: any non-zero vector $(m, n) \in \mathbb{Z}^2$ can be written uniquely as $r(c, d)$ with r (the greatest common divisor of m and n) a positive integer and c and d coprime integers, so

$$G_k(z) \;=\; \zeta(k)\, E_k(z)\,, \tag{11}$$

where $\zeta(k) = \sum_{r\ge1} 1/r^k$ is the value at k of the Riemann zeta function. It may therefore seem pointless to have introduced both definitions. But in fact, this is not the case. First of all, each definition gives a distinct point of view and has advantages in certain settings which are encountered at later points in the theory: the E_k definition is better in contexts like the famous Rankin-Selberg method where one integrates the product of the Eisenstein series with another modular form over a fundamental domain, while the G_k definition is better for analytic calculations and for the Fourier development given in §2.2.

Moreover, if one passes to other groups, then there are σ Eisenstein series of each type, where σ is the number of cusps, and, although they span the same vector space, they are not individually proportional. In fact, we will actually want to introduce a *third* normalization

$$\mathbb{G}_k(z) \;=\; \frac{(k-1)!}{(2\pi i)^k}\, G_k(z) \tag{12}$$

because, as we will see below, it has Fourier coefficients which are rational numbers (and even, with one exception, integers) and because it is a normalized eigenfunction for the Hecke operators discussed in §4.

As a first application, we can now determine the ring structure of $M_*(\Gamma_1)$

Proposition 4. *The ring $M_*(\Gamma_1)$ is freely generated by the modular forms E_4 and E_6.*

Corollary. *The inequality* (7) *for the dimension of $M_k(\Gamma_1)$ is an equality for all even $k \geq 0$.*

Proof. The essential point is to show that the modular forms $E_4(z)$ and $E_6(z)$ are algebraically independent. To see this, we first note that the forms $E_4(z)^3$ and $E_6(z)^2$ of weight 12 cannot be proportional. Indeed, if we had $E_6(z)^2 = \lambda E_4(z)^3$ for some (necessarily non-zero) constant λ, then the meromorphic modular form $f(z) = E_6(z)/E_4(z)$ of weight 2 would satisfy $f^2 = \lambda E_4$ (and also $f^3 = \lambda^{-1} E_6$) and would hence be holomorphic (a function whose square is holomorphic cannot have poles), contradicting the inequality $\dim M_2(\Gamma_1) \leq 0$ of Corollary 1 of Proposition 2. But *any* two modular forms f_1 and f_2 of the same weight which are not proportional are necessarily algebraically independent. Indeed, if $P(X,Y)$ is any polynomial in $\mathbb{C}[X,Y]$ such that $P(f_1(z), f_2(z)) \equiv 0$, then by considering the weights we see that $P_d(f_1, f_2)$ has to vanish identically for each homogeneous component P_d of P. But $P_d(f_1,f_2)/f_2^d = p(f_1/f_2)$ for some polynomial $p(t)$ in one variable, and since p has only finitely many roots we can only have $P_d(f_1,f_2) \equiv 0$ if f_1/f_2 is a constant. It follows that E_4^3 and E_6^2, and hence also E_4 and E_6, are algebraically independent. But then an easy calculation shows that the dimension of the weight k part of the subring of $M_*(\Gamma_1)$ which they generate equals the right-hand side of the inequality (7), so that the proposition and corollary follow from this inequality.

2.2 Fourier Expansions of Eisenstein Series

Recall from (3) that any modular form on Γ_1 has a Fourier expansion of the form $\sum_{n=0}^{\infty} a_n q^n$, where $q = e^{2\pi i z}$. The coefficients a_n often contain interesting arithmetic information, and it is this that makes modular forms important for classical number theory. For the Eisenstein series, normalized by (12), the coefficients are given by:

Proposition 5. *The Fourier expansion of the Eisenstein series* $\mathbb{G}_k(z)$ *(k even, $k>2$) is*

$$\mathbb{G}_k(z) \;=\; -\frac{B_k}{2k} \,+\, \sum_{n=1}^{\infty} \sigma_{k-1}(n)\,q^n\,, \tag{13}$$

where B_k *is the* k*th Bernoulli number and where* $\sigma_{k-1}(n)$ *for* $n \in \mathbb{N}$ *denotes the sum of the* $(k-1)$*st powers of the positive divisors of* n.

We recall that the Bernoulli numbers are defined by the generating function $\sum_{k=0}^{\infty} B_k x^k/k! = x/(e^x-1)$ and that the first values of B_k ($k>0$ even) are given by $B_2=\frac{1}{6}$, $B_4=-\frac{1}{30}$, $B_6=\frac{1}{42}$, $B_8=-\frac{1}{30}$, $B_{10}=\frac{5}{66}$, $B_{12}=-\frac{691}{2730}$, and $B_{14}=\frac{7}{6}$.

Proof. A well known and easily proved identity of Euler states that

$$\sum_{n\in\mathbb{Z}} \frac{1}{z+n} \;=\; \frac{\pi}{\tan \pi z} \qquad \bigl(\, z \in \mathbb{C}\setminus\mathbb{Z}\,\bigr)\,, \tag{14}$$

where the sum on the left, which is not absolutely convergent, is to be interpreted as a Cauchy principal value ($=\lim \sum_{-M}^{N}$ where M, N tend to infinity with $M-N$ bounded). The function on the right is periodic of period 1 and its Fourier expansion for $z \in \mathfrak{H}$ is given by

$$\frac{\pi}{\tan \pi z} \;=\; \pi\,\frac{\cos \pi z}{\sin \pi z} \;=\; \pi i\,\frac{e^{\pi i z}+e^{-\pi i z}}{e^{\pi i z}-e^{-\pi i z}} \;=\; -\pi i\,\frac{1+q}{1-q} \;=\; -2\pi i\left(\frac{1}{2}+\sum_{r=1}^{\infty} q^r\right),$$

where $q=e^{2\pi i z}$. Substitute this into (14), differentiate $k-1$ times and divide by $(-1)^{k-1}(k-1)!$ to get

$$\sum_{n\in\mathbb{Z}} \frac{1}{(z+n)^k} \;=\; \frac{(-1)^{k-1}}{(k-1)!}\,\frac{d^{k-1}}{dz^{k-1}}\left(\frac{\pi}{\tan \pi z}\right) \;=\; \frac{(-2\pi i)^k}{(k-1)!}\,\sum_{r=1}^{\infty} r^{k-1}\,q^r \qquad (k\ge 2,\; z\in\mathfrak{H})\,,$$

an identity known as Lipschitz's formula. Now the Fourier expansion of G_k ($k>2$ even) is obtained immediately by splitting up the sum in (10) into the terms with $m=0$ and those with $m\neq 0$:

$$\begin{aligned} G_k(z) \;&=\; \frac{1}{2}\sum_{\substack{n\in\mathbb{Z}\\ n\neq 0}} \frac{1}{n^k} \,+\, \frac{1}{2}\sum_{\substack{m,\,n\in\mathbb{Z}\\ m\neq 0}} \frac{1}{(mz+n)^k} \;=\; \sum_{n=1}^{\infty}\frac{1}{n^k} + \sum_{m=1}^{\infty}\sum_{n=-\infty}^{\infty} \frac{1}{(mz+n)^k} \\ &=\; \zeta(k) \,+\, \frac{(2\pi i)^k}{(k-1)!}\,\sum_{m=1}^{\infty}\sum_{r=1}^{\infty} r^{k-1}\,q^{mr} \\ &=\; \frac{(2\pi i)^k}{(k-1)!}\left(-\frac{B_k}{2k} \,+\, \sum_{n=1}^{\infty}\sigma_{k-1}(n)\,q^n\right), \end{aligned}$$

where in the last line we have used Euler's evaluation of $\zeta(k)$ ($k>0$ even) in terms of Bernoulli numbers. The result follows.

The first three examples of Proposition 5 are the expansions

$$\begin{aligned}\mathbb{G}_4(z) &= \frac{1}{240} + q + 9q^2 + 28q^3 + 73q^4 + 126q^5 + 252q^6 + \cdots,\\ \mathbb{G}_6(z) &= -\frac{1}{504} + q + 33q^2 + 244q^3 + 1057q^4 + \cdots,\\ \mathbb{G}_8(z) &= \frac{1}{480} + q + 129q^2 + 2188q^3 + \cdots.\end{aligned}$$

The other two normalizations of these functions are given by

$$\begin{aligned}G_4(z) &= \frac{16\,\pi^4}{3!}\,\mathbb{G}_4(z) = \frac{\pi^4}{90}\,E_4(z), \qquad E_4(z) = 1 + 240q + 2160q^2 + \cdots,\\ G_6(z) &= -\frac{64\,\pi^6}{5!}\,\mathbb{G}_6(z) = \frac{\pi^6}{945}\,E_6(z), \quad E_6(z) = 1 - 504q - 16632q^2 - \cdots,\\ G_8(z) &= \frac{256\,\pi^8}{7!}\,\mathbb{G}_8(z) = \frac{\pi^8}{9450}\,E_8(z), \quad E_8(z) = 1 + 480q + 61920q^2 + \cdots.\end{aligned}$$

Remark. We have discussed only Eisenstein series on the full modular group in detail, but there are also various kinds of Eisenstein series for subgroups $\Gamma \subset \Gamma_1$. We give one example. Recall that a *Dirichlet character* modulo $N \in \mathbb{N}$ is a homomorphism $\chi : (\mathbb{Z}/N\mathbb{Z})^* \to \mathbb{C}^*$, extended to a map $\chi : \mathbb{Z} \to \mathbb{C}$ (traditionally denoted by the same letter) by setting $\chi(n)$ equal to $\chi(n \bmod N)$ if $(n, N) = 1$ and to 0 otherwise. If χ is a non-trivial Dirichlet character and k a positive integer with $\chi(-1) = (-1)^k$, then there is an Eisenstein series having the Fourier expansion

$$\mathbb{G}_{k,\chi}(z) \;=\; c_k(\chi) \;+\; \sum_{n=1}^{\infty}\Bigl(\sum_{d|n}\chi(d)\,d^{k-1}\Bigr)\,q^n$$

which is a "modular form of weight k and character χ on $\Gamma_0(N)$." (This means that $\mathbb{G}_{k,\chi}(\frac{az+b}{cz+d}) = \chi(a)(cz+d)^k\mathbb{G}_{k,\chi}(z)$ for any $z \in \mathfrak{H}$ and any $\left(\begin{smallmatrix} a & b \\ c & d\end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{Z})$ with $c \equiv 0 \pmod{N}$.) Here $c_k(\chi) \in \overline{\mathbb{Q}}$ is a suitable constant, given explicitly by $c_k(\chi) = \frac{1}{2}L(1-k,\chi)$, where $L(s,\chi)$ is the analytic continuation of the Dirichlet series $\sum_{n=1}^{\infty}\chi(n)n^{-s}$.

The simplest example, for $N = 4$ and $\chi = \chi_{-4}$ the Dirichlet character modulo 4 given by

$$\chi_{-4}(n) \;=\; \begin{cases} +1 & \text{if } n \equiv 1 \pmod{4},\\ -1 & \text{if } n \equiv 3 \pmod{4},\\ \;\;0 & \text{if } n \text{ is even}\end{cases} \tag{15}$$

and $k = 1$, is the series

$$\mathbb{G}_{1,\chi_{-4}}(z) = c_1(\chi_{-4}) + \sum_{n=1}^{\infty}\Biggl(\sum_{d|n}\chi_{-4}(d)\Biggr)\,q^n = \frac{1}{4} + q + q^2 + q^4 + 2q^5 + q^8 + \cdots. \tag{16}$$

(The fact that $L(0,\chi_{-4}) = 2c_1(\chi_{-4}) = \dfrac{1}{2}$ is equivalent via the functional equation of $L(s,\chi_{-4})$ to Leibnitz's famous formula $L(1,\chi_{-4}) = 1 - \dfrac{1}{3} + \dfrac{1}{5} - \cdots = \dfrac{\pi}{4}$.) We will see this function again in §3.1.

♠ Identities Involving Sums of Powers of Divisors

We now have our first explicit examples of modular forms and their Fourier expansions and can immediately deduce non-trivial number-theoretic identities. For instance, each of the spaces $M_4(\Gamma_1)$, $M_6(\Gamma_1)$, $M_8(\Gamma_1)$, $M_{10}(\Gamma_1)$ and $M_{14}(\Gamma_1)$ has dimension exactly 1 by the corollary to Proposition 2, and is therefore spanned by the Eisenstein series $E_k(z)$ with leading coefficient 1, so we immediately get the identities

$$\begin{aligned} E_4(z)^2 \;=\; E_8(z)\,, \quad E_4(z)E_6(z) \;=\; E_{10}(z)\,, \\ E_6(z)E_8(z) \;=\; E_4(z)E_{10}(z) \;=\; E_{14}(z)\,. \end{aligned}$$

Each of these can be combined with the Fourier expansion given in Proposition 5 to give an identity involving the sums-of-powers-of-divisors functions $\sigma_{k-1}(n)$, the first and the last of these being

$$\begin{aligned} \sum_{m=1}^{n-1} \sigma_3(m)\sigma_3(n-m) &= \frac{\sigma_7(n) - \sigma_3(n)}{120}\,, \\ \sum_{m=1}^{n-1} \sigma_3(m)\sigma_9(n-m) &= \frac{\sigma_{13}(n) - 11\sigma_9(n) + 10\sigma_3(n)}{2640}\,. \end{aligned}$$

Of course similar identities can be obtained from modular forms in higher weights, even though the dimension of $M_k(\Gamma_1)$ is no longer equal to 1. For instance, the fact that $M_{12}(\Gamma_1)$ is 2-dimensional and contains the three modular forms E_4E_8, E_6^2 and E_{12} implies that the three functions are linearly dependent, and by looking at the first two terms of the Fourier expansions we find that the relation between them is given by $441E_4E_8 + 250E_6^2 = 691E_{12}$, a formula which the reader can write out explicitly as an identity among sums-of-powers-of-divisors functions if he or she is so inclined. It is not easy to obtain any of these identities by direct number-theoretical reasoning (although in fact it can be done). ♡

2.3 The Eisenstein Series of Weight 2

In §2.1 and §2.2 we restricted ourselves to the case when $k > 2$, since then the series (9) and (10) are absolutely convergent and therefore define modular forms of weight k. But the final formula (13) for the Fourier expansion of $\mathbb{G}_k(z)$ converges rapidly and defines a holomorphic function of z also for $k = 2$, so in this weight we can simply *define* the Eisenstein series $\mathbb{G}_2$, G_2 and E_2 by equations (13), (12), and (11), respectively, i.e.,

$$\begin{aligned}\mathbb{G}_2(z) &= -\frac{1}{24} + \sum_{n=1}^{\infty} \sigma_1(n)\, q^n \;=\; -\frac{1}{24} + q + 3q^2 + 4q^3 + 7q^4 + 6q^5 + \cdots\,, \\ G_2(z) &= -4\pi^2\, \mathbb{G}_2(z)\,, \quad E_2(z) \;=\; \frac{6}{\pi^2}\, G_2(z) \;=\; 1 - 24q - 72q^2 - \cdots\,. \end{aligned} \tag{17}$$

Moreover, the same proof as for Proposition 5 still shows that $G_2(z)$ is given by the expression (10), if we agree to carry out the summation over n first and then over m:

$$G_2(z) \;=\; \frac{1}{2}\sum_{n\neq 0}\frac{1}{n^2} + \frac{1}{2}\sum_{m\neq 0}\sum_{n\in\mathbb{Z}}\frac{1}{(mz+n)^2}\,. \tag{18}$$

The only difference is that, because of the non-absolute convergence of the double series, we can no longer interchange the order of summation to get the modular transformation equation $G_2(-1/z) = z^2 G_2(z)$. (The equation $G_2(z+1) = G_2(z)$, of course, still holds just as for higher weights.) Nevertheless, the function $G_2(z)$ and its multiples $E_2(z)$ and $\mathbb{G}_2(z)$ do have some modular properties and, as we will see later, these are important for many applications.

Proposition 6. *For $z \in \mathfrak{H}$ and $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL(2,\mathbb{Z})$ we have*

$$G_2\left(\frac{az+b}{cz+d}\right) \;=\; (cz+d)^2\, G_2(z) \;-\; \pi i c(cz+d)\,. \tag{19}$$

Proof. There are many ways to prove this. We sketch one, due to Hecke, since the method is useful in many other situations. The series (10) for $k = 2$ does not converge absolutely, but it is just at the edge of convergence, since $\sum_{m,n} |mz+n|^{-\lambda}$ converges for any real number $\lambda > 2$. We therefore modify the sum slightly by introducing

$$G_{2,\varepsilon}(z) \;=\; \frac{1}{2}\sum_{m,\,n}{}^{'}\frac{1}{(mz+n)^2\,|mz+n|^{2\varepsilon}} \qquad (z \in \mathfrak{H},\ \varepsilon > 0)\,. \tag{20}$$

(Here $\sum'$ means that the value $(m,n) = (0,0)$ is to be omitted from the summation.) The new series converges absolutely and transforms by $G_{2,\varepsilon}\bigl(\frac{az+b}{cz+d}\bigr) = (cz+d)^2|cz+d|^{2\varepsilon} G_{2,\varepsilon}(z)$. We claim that $\lim\limits_{\varepsilon\to 0} G_{2,\varepsilon}(z)$ exists and equals $G_2(z) - \pi/2y$, where $y = \Im(z)$. It follows that each of the three non-holomorphic functions

$$G_2^*(z) \;=\; G_2(z) - \frac{\pi}{2y}\,, \quad E_2^*(z) \;=\; E_2(z) - \frac{3}{\pi y}\,, \quad \mathbb{G}_2^*(z) \;=\; \mathbb{G}_2(z) + \frac{1}{8\pi y} \tag{21}$$

transforms like a modular form of weight 2, and from this one easily deduces the transformation equation (19) and its analogues for E_2 and $\mathbb{G}_2$. To prove the claim, we define a function I_ε by

$$I_\varepsilon(z) \;=\; \int_{-\infty}^{\infty} \frac{dt}{(z+t)^2\,|z+t|^{2\varepsilon}} \qquad \bigl(z \in \mathfrak{H},\ \varepsilon > -\tfrac{1}{2}\bigr)\,.$$

Then for $\varepsilon > 0$ we can write

$$\begin{aligned} G_{2,\varepsilon} - \sum_{m=1}^{\infty} I_\varepsilon(mz) \;&=\; \sum_{n=1}^{\infty} \frac{1}{n^{2+2\varepsilon}} \\ &+ \sum_{m=1}^{\infty}\sum_{n=-\infty}^{\infty}\biggl[\frac{1}{(mz+n)^2\,|mz+n|^{2\varepsilon}} - \int_n^{n+1}\frac{dt}{(mz+t)^2|mz+t|^{2\varepsilon}}\biggr]\,. \end{aligned}$$

Both sums on the right converge absolutely and locally uniformly for $\varepsilon > -\frac{1}{2}$ (the second one because the expression in square brackets is $\mathrm{O}\bigl(|mz{+}n|^{-3-2\varepsilon}\bigr)$ by the mean-value theorem, which tells us that $f(t) - f(n)$ for any differentiable function f is bounded in $n \le t \le n+1$ by $\max_{n\le u\le n+1}|f'(u)|$), so the limit of the expression on the right as $\varepsilon \to 0$ exists and can be obtained simply by putting $\varepsilon = 0$ in each term, where it reduces to $G_2(z)$ by (18). On the other hand, for $\varepsilon > -\frac{1}{2}$ we have

$$\begin{aligned} I_\varepsilon(x+iy) \;&=\; \int_{-\infty}^{\infty}\frac{dt}{(x+t+iy)^2\,((x+t)^2+y^2)^\varepsilon} \\ &=\; \int_{-\infty}^{\infty}\frac{dt}{(t+iy)^2\,(t^2+y^2)^\varepsilon} \;=\; \frac{I(\varepsilon)}{y^{1+2\varepsilon}}\,, \end{aligned}$$

where $I(\varepsilon) = \int_{-\infty}^{\infty}(t{+}i)^{-2}(t^2{+}1)^{-\varepsilon}dt$, so $\sum_{m=1}^{\infty} I_\varepsilon(mz) = I(\varepsilon)\zeta(1{+}2\varepsilon)/y^{1+2\varepsilon}$ for $\varepsilon > 0$. Finally, we have $I(0) = 0$ (obvious),

$$I'(0) \;=\; -\int_{-\infty}^{\infty}\frac{\log(t^2+1)}{(t+i)^2}\,dt \;=\; \biggl(\frac{1+\log(t^2+1)}{t+i} \;-\; \tan^{-1}t\biggr)\bigg|_{-\infty}^{\infty} \;=\; -\,\pi\,,$$

and $\zeta(1{+}2\varepsilon) = \dfrac{1}{2\varepsilon} + \mathrm{O}(1)$, so the product $I(\varepsilon)\zeta(1{+}2\varepsilon)/y^{1+2\varepsilon}$ tends to $-\pi/2y$ as $\varepsilon \to 0$. The claim follows.

Remark. The transformation equation (18) says that G_2 is an example of what is called a *quasimodular* form, while the functions G_2^*, E_2^* and $\mathbb{G}_2^*$ defined in (21) are so-called *almost holomorphic modular forms* of weight 2. We will return to this topic in Section 5.

2.4 The Discriminant Function and Cusp Forms

For $z \in \mathfrak{H}$ we define the *discriminant function* $\Delta(z)$ by the formula

$$\Delta(z) \;=\; e^{2\pi iz}\prod_{n=1}^{\infty}\bigl(1 - e^{2\pi inz}\bigr)^{24}\,. \tag{22}$$

(The name comes from the connection with the discriminant of the elliptic curve $E_z = \mathbb{C}/(\mathbb{Z}.z + \mathbb{Z}.1)$, but we will not discuss this here.) Since $|e^{2\pi iz}| < 1$ for $z \in \mathfrak{H}$, the terms of the infinite product are all non-zero and tend exponentially rapidly to 1, so the product converges everywhere and defines a holomorphic and everywhere non-zero function in the upper half-plane. This function turns out to be a modular form and plays a special role in the entire theory.

Proposition 7. *The function* $\Delta(z)$ *is a modular form of weight* 12 *on* $SL(2,\mathbb{Z})$*.*

Proof. Since $\Delta(z) \neq 0$, we can consider its logarithmic derivative. We find

$$\frac{1}{2\pi i}\,\frac{d}{dz}\log\Delta(z) \;=\; 1-24\sum_{n=1}^{\infty}\frac{n\,e^{2\pi inz}}{1-e^{2\pi inz}} \;=\; 1-24\sum_{m=1}^{\infty}\sigma_1(m)\,e^{2\pi imz} \;=\; E_2(z)\,,$$

where the second equality follows by expanding $\dfrac{e^{2\pi inz}}{1-e^{2\pi inz}}$ as a geometric series $\sum_{r=1}^{\infty} e^{2\pi irnz}$ and interchanging the order of summation, and the third equality from the definition of $E_2(z)$ in (17). Now from the transformation equation for E_2 (obtained by comparing (19) and(11)) we find

$$\frac{1}{2\pi i}\,\frac{d}{dz}\log\biggl(\frac{\Delta\bigl(\frac{az+b}{cz+d}\bigr)}{(cz+d)^{12}\Delta(z)}\biggr) = \frac{1}{(cz+d)^2}\,E_2\biggl(\frac{az+b}{cz+d}\biggr) - \frac{12}{2\pi i}\,\frac{c}{cz+d} - E_2(z)$$
$$= 0\,.$$

In other words, $(\Delta|_{12}\gamma)(z) = C(\gamma)\,\Delta(z)$ for all $z \in \mathfrak{H}$ and all $\gamma \in \Gamma_1$, where $C(\gamma)$ is a non-zero complex number depending only on γ, and where $\Delta|_{12}\gamma$ is defined as in (8). It remains to show that $C(\gamma) = 1$ for all γ. But $C : \Gamma_1 \to \mathbb{C}^*$ is a homomorphism because $\Delta \mapsto \Delta|_{12}\gamma$ is a group action, so it suffices to check this for the generators $T = \left(\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right)$ and $S = \left(\begin{smallmatrix}0&-1\\1&0\end{smallmatrix}\right)$ of Γ_1. The first is obvious since $\Delta(z)$ is a power series in $e^{2\pi iz}$ and hence periodic of period 1, while the second follows by substituting $z = i$ into the equation $\Delta(-1/z) = C(S)\,z^{12}\Delta(z)$ and noting that $\Delta(i) \neq 0$.

Let us look at this function $\Delta(z)$ more carefully. We know from Corollary 1 to Proposition 2 that the space $M_{12}(\Gamma_1)$ has dimension at most 2, so $\Delta(z)$ must be a linear combination of the two functions $E_4(z)^3$ and $E_6(z)^2$. From the Fourier expansions $E_4^3 = 1 + 720q + \cdots$, $E_6(z)^2 = 1 - 1008q + \cdots$ and $\Delta(z) = q + \cdots$ we see that this relation is given by

$$\Delta(z) \;=\; \frac{1}{1728}\left(E_4(z)^3 \;-\; E_6(z)^2\right). \tag{23}$$

This identity permits us to give another, more explicit, version of the fact that every modular form on Γ_1 is a polynomial in E_4 and E_6 (Proposition 4). Indeed, let $f(z)$ be a modular form of arbitrary even weight $k \geq 4$, with Fourier expansion as in (3). Choose integers a, $b \geq 0$ with $4a + 6b = k$ (this is always possible) and set $h(z) = \bigl(f(z) - a_0 E_4(z)^a E_6(z)^b\bigr)/\Delta(z)$. This function is holomorphic in $\mathfrak{H}$ (because $\Delta(z) \neq 0$) and also at infinity (because $f - a_0 E_4^a E_6^b$ has a Fourier expansion with no constant term and the Fourier expansion of Δ begins with q), so it is a modular form of weight $k - 12$. By induction on the weight, h is a polynomial in E_4 and E_6, and then from $f = a_0 E_4^a E_6^b + \Delta h$ and (23) we see that f also is.

In the above argument we used that $\Delta(z)$ has a Fourier expansion beginning $q + \mathrm{O}(q^2)$ and that $\Delta(z)$ is never zero in the upper half-plane. We deduced both facts from the product expansion (22), but it is perhaps worth noting that this is not necessary: if we were simply to *define* $\Delta(z)$ by equation (23), then the fact that its Fourier expansion begins with q would follow from the knowledge of the first two Fourier coefficients of E_4 and E_6, and the fact that it never vanishes in $\mathfrak{H}$ would then follow from Proposition 2 because the total number $k/12 = 1$ of Γ_1-inequivalent zeros of Δ is completely accounted for by the first-order zero at infinity.

We can now make the concrete normalization of the isomorphism between $\overline{\Gamma_1\backslash\mathfrak{H}}$ and $\mathbb{P}^1(\mathbb{C})$ mentioned after Corollary 2 of Proposition 2. In the notation of that proposition, choose $f(z) = E_4(z)^3$ and $g(z) = \Delta(z)$. Their quotient is then the modular function

$$j(z) \;=\; \frac{E_4(z)^3}{\Delta(z)} \;=\; q^{-1} \,+\, 744 \,+\, 196884\,q \,+\, 21493760\,q^2 \,+\, \cdots\,,$$

called the *modular invariant*. Since $\Delta(z) \neq 0$ for $z \in \mathfrak{H}$, this function is finite in $\mathfrak{H}$ and defines an isomorphism from $\Gamma_1\backslash\mathfrak{H}$ to $\mathbb{C}$ as well as from $\overline{\Gamma_1\backslash\mathfrak{H}}$ to $\mathbb{P}^1(\mathbb{C})$.

The next (and most interesting) remarks about $\Delta(z)$ concern its Fourier expansion. By multiplying out the product in (22) we obtain the expansion

$$\Delta(z) \;=\; q\prod_{n=1}^{\infty}\bigl(1 - q^n\bigr)^{24} \;=\; \sum_{n=1}^{\infty}\tau(n)\,q^n \tag{24}$$

where $q = e^{2\pi i z}$ as usual (this is the last time we will repeat this!) and the coefficients $\tau(n)$ are certain integers, the first values being given by the table

n	1	2	3	4	5	6	7	8	9	10
$\tau(n)$	1	-24	252	-1472	4830	-6048	-16744	84480	-113643	-115920

Ramanujan calculated the first 30 values of $\tau(n)$ in 1915 and observed several remarkable properties, notably the multiplicativity property that $\tau(pq) = \tau(p)\tau(q)$ if p and q are distinct primes (e.g., $-6048 = -24 \cdot 252$ for $p = 2$, $q = 3$) and $\tau(p^2) = \tau(p)^2 - p^{11}$ if p is prime (e.g., $-1472 = (-24)^2 - 2048$ for $p = 2$). This was proved by Mordell the next year and later generalized by Hecke to the theory of Hecke operators, which we will discuss in §4.

Ramanujan also observed that $|\tau(p)|$ was bounded by $2p^5\sqrt{p}$ for primes $p < 30$, and conjectured that this holds for all p. This property turned out to be immeasurably deeper than the assertion about multiplicativity and was only proved in 1974 by Deligne as a consequence of his proof of the famous Weil conjectures (and of his previous, also very deep, proof that these conjectures implied Ramanujan's). However, the weaker inequality $|\tau(p)| \le Cp^6$ with some effective constant $C > 0$ is much easier and was proved in the 1930's by Hecke. We reproduce Hecke's proof, since it is simple. In fact, the proof applies to a much more general class of modular forms. Let us call a modular form on Γ_1 a *cusp form* if the constant term a_0 in the Fourier expansion (3) is zero. Since the constant term of the Eisenstein series $\mathbb{G}_k(z)$ is non-zero, any modular form can be written uniquely as a linear combination of an Eisenstein series and a cusp form of the same weight. For the former the Fourier coefficients are given by (13) and grow like n^{k-1} (since $n^{k-1} \le \sigma_{k-1}(n) < \zeta(k-1)n^{k-1}$). For the latter, we have:

Proposition 8. *Let $f(z)$ be a cusp form of weight k on Γ_1 with Fourier expansion $\sum_{n=1}^{\infty} a_n q^n$. Then $|a_n| \le Cn^{k/2}$ for all n, for some constant C depending only on f.*

Proof. From equations (1) and (2) we see that the function $z \mapsto y^{k/2}|f(z)|$ on $\mathfrak{H}$ is Γ_1-invariant. This function tends rapidly to 0 as $y = \Im(z) \to \infty$ (because $f(z) = \mathrm{O}(q)$ by assumption and $|q| = e^{-2\pi y}$), so from the form of the fundamental domain of Γ_1 as given in Proposition 1 it is clearly bounded. Thus we have the estimate

$$|f(z)| \;\le\; c\,y^{-k/2} \qquad (z \;=\; x+iy \in \mathfrak{H}) \tag{25}$$

for some $c > 0$ depending only on f. Now the integral representation

$$a_n \;=\; e^{2\pi n y} \int_0^1 f(x+iy)\, e^{-2\pi i n x}\, dx$$

for a_n, valid for any $y > 0$, show that $|a_n| \le cy^{-k/2}e^{2\pi n y}$. Taking $y = 1/n$ (or, optimally, $y = k/4\pi n$) gives the estimate of the proposition with $C = c\,e^{2\pi}$ (or, optimally, $C = c\,(4\pi e/k)^{k/2}$).

Remark. The definition of cusp forms given above is actually valid only for the full modular group Γ_1 or for other groups having only one cusp. In general one must require the vanishing of the constant term of the Fourier expansion of f, suitably defined, at every cusp of the group Γ, in which case it again follows that f can be estimated as in (25). Actually, it is easier to simply *define* cusp forms of weight k as modular forms for which $y^{k/2}f(x+iy)$ is bounded, a definition which is equivalent but does not require the explicit knowledge of the Fourier expansion of the form at every cusp.

♠ Congruences for $\tau(n)$

As a mini-application of the calculations of this and the preceding sections we prove two simple congruences for the Ramanujan tau-function defined by equation (24). First of all, let us check directly that the coefficient $\tau(n)$ of q^n of the function defined by (23) is integral for all n. (This fact is, of course, obvious from equation (22).) We have

$$\Delta \;=\; \frac{(1+240A)^3-(1-504B)^2}{1728} \;=\; 5\,\frac{A-B}{12}+B+100A^2-147B^2+8000A^3 \tag{26}$$

with $A=\sum_{n=1}^{\infty}\sigma_3(n)q^n$ and $B=\sum_{n=1}^{\infty}\sigma_5(n)q^n$. But $\sigma_5(n)-\sigma_3(n)$ is divisible by 12 for every n (because 12 divides d^5-d^3 for every d), so $(A-B)/12$ has integral coefficients. This gives the integrality of $\tau(n)$, and even a congruence modulo 2. Indeed, we actually have $\sigma_5(n)\equiv\sigma_3(n)\pmod{24}$, because $d^3(d^2-1)$ is divisible by 24 for every d, so $(A-B)/12$ has even coefficients and (26) gives $\Delta\equiv B+B^2\pmod 2$ or, recalling that $(\sum a_nq^n)^2\equiv\sum a_nq^{2n}\pmod 2$ for every power series $\sum a_nq^n$ with integral coefficients, $\tau(n)\equiv\sigma_5(n)+\sigma_5(n/2)\pmod 2$, where $\sigma_5(n/2)$ is defined as 0 if $2\nmid n$. But $\sigma_5(n)$, for any integer n, is congruent modulo 2 to the sum of the odd divisors of n, and this is odd if and only if n is a square or twice a square, as one sees by writing $n=2^sn_0$ with n_0 odd and pairing the complementary divisors of n_0. It follows that $\sigma_5(n)+\sigma_5(n/2)$ is odd if and only if n is an odd square, so we get the congruence:

$$\tau(n)\;\equiv\;\begin{cases}1\pmod 2 & \text{if } n \text{ is an odd square},\\ 0\pmod 2 & \text{otherwise}.\end{cases} \tag{27}$$

In a different direction, from $\dim M_{12}(\Gamma_1)=2$ we immediately deduce the linear relation

$$\mathbb{G}_{12}(z)\;=\;\Delta(z)+\frac{691}{156}\left(\frac{E_4(z)^3}{720}+\frac{E_6(z)^3}{1008}\right)$$

and from this a famous congruence of Ramanujan,

$$\tau(n)\;\equiv\;\sigma_{11}(n)\pmod{691}\qquad\bigl(\forall n\geq 1\bigr), \tag{28}$$

where the “691” comes from the numerator of the constant term $-B_{12}/24$ of $\mathbb{G}_{12}$. ♡

3 Theta Series

If Q is a positive definite integer-valued quadratic form in m variables, then there is an associated modular form of weight $m/2$, called the *theta series* of Q, whose nth Fourier coefficient for every integer $n\geq 0$ is the number of representations of n by Q. This provides at the same time one of the main constructions of modular forms and one of the most important sources of applications of the theory. In 3.1 we consider unary theta series ($m=1$), while the general case is discussed in 3.2. The unary case is the most classical, going back to Jacobi, and already has many applications. It is also the basis of the general theory, because any quadratic form can be diagonalized over $\mathbb{Q}$ (i.e., by passing to a suitable sublattice it becomes the direct sum of m quadratic forms in one variable).

3.1 Jacobi's Theta Series

The simplest theta series, corresponding to the unary (one-variable) quadratic form $x \mapsto x^2$, is Jacobi's theta function

$$\theta(z) \;=\; \sum_{n\in\mathbb{Z}} q^{n^2} \;=\; 1 \,+\, 2q \,+\, 2q^4 \,+\, 2q^9 \,+\, \cdots\,, \tag{29}$$

where $z \in \mathfrak{H}$ and $q = e^{2\pi i z}$ as usual. Its modular transformation properties are given as follows.

Proposition 9. *The function $\theta(z)$ satisfies the two functional equations*

$$\theta(z+1) \;=\; \theta(z)\,, \qquad \theta\left(\frac{-1}{4z}\right) \;=\; \sqrt{\frac{2z}{i}}\,\theta(z) \qquad (z \in \mathfrak{H})\,. \tag{30}$$

Proof. The first equation in (30) is obvious since $\theta(z)$ depends only on q. For the second, we use the Poisson transformation formula. Recall that this formula says that for any function $f : \mathbb{R} \to \mathbb{C}$ which is smooth and small at infinity, we have $\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \widetilde{f}(n)$, where $\widetilde{f}(y) = \int_{-\infty}^{\infty} e^{2\pi i x y}\, f(x)\, dx$ is the Fourier transform of f. (*Proof*: the sum $\sum_{n\in\mathbb{Z}} f(n+x)$ is convergent and defines a function $g(x)$ which is periodic of period 1 and hence has a Fourier expansion $g(x) = \sum_{n\in\mathbb{Z}} c_n e^{2\pi i n x}$ with $c_n = \int_0^1 g(x) e^{-2\pi i n x} dx = \widetilde{f}(-n)$, so $\sum_n f(n) = g(0) = \sum_n c_n = \sum_n \widetilde{f}(-n) = \sum_n \widetilde{f}(n)$.) Applying this to the function $f(x) = e^{-\pi t x^2}$, where t is a positive real number, and noting that

$$\widetilde{f}(y) \;=\; \int_{-\infty}^{\infty} e^{-\pi t x^2 + 2\pi i x y}\, dx \;=\; \frac{e^{-\pi y^2/t}}{\sqrt{t}} \int_{-\infty}^{\infty} e^{-\pi u^2}\, du \;=\; \frac{e^{-\pi y^2/t}}{\sqrt{t}}$$

(substitution $u = \sqrt{t}\,(x - iy/t)$ followed by a shift of the path of integration), we obtain

$$\sum_{n=-\infty}^{\infty} e^{-\pi n^2 t} \;=\; \frac{1}{\sqrt{t}} \sum_{n=-\infty}^{\infty} e^{-\pi n^2/t} \qquad (t > 0)\,.$$

This proves the second equation in (30) for $z = it/2$ lying on the positive imaginary axis, and the general case then follows by analytic continuation.

The point is now that the two transformations $z \mapsto z+1$ and $z \mapsto -1/4z$ generate a subgroup of $\mathrm{SL}(2,\mathbb{R})$ which is commensurable with $\mathrm{SL}(2,\mathbb{Z})$, so (30) implies that the function $\theta(z)$ is a modular form of weight $1/2$. (We have not defined modular forms of half-integral weight and will not discuss their theory in these notes, but the reader can simply interpret this statement as saying that $\theta(z)^2$ is a modular form of weight 1.) More specifically, for every $N \in \mathbb{N}$ we have the "congruence subgroup" $\Gamma_0(N) \subseteq \Gamma_1 = \mathrm{SL}(2,\mathbb{Z})$, consisting of matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_1$ with c divisible by N, and the larger group $\Gamma_0^+(N) = \langle \Gamma_0(N), W_N \rangle = \Gamma_0(N) \cup \Gamma_0(N)W_N$, where $W_N = \frac{1}{\sqrt{N}}\left(\begin{smallmatrix} 0 & -1 \\ N & 0 \end{smallmatrix}\right)$ ("Fricke involution") is an element of $\mathrm{SL}(2,\mathbb{R})$ of order 2 which normalizes $\Gamma_0(N)$. The group $\Gamma_0^+(N)$ contains the elements $T = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and W_N for any N. In general they generate a subgroup of infinite index, so that to check the modularity of a given function it does not suffice to verify its behavior just for $z \mapsto z+1$ and $z \mapsto -1/Nz$, but for $N = 4$ (like for $N = 1$!) they generate the full group and this *is* sufficient. The proof is simple. Since $W_N^2 = -1$, it is sufficient to show that the two matrices T and $\widetilde{T} = W_4 T W_4^{-1} = \left(\begin{smallmatrix} 1 & 0 \\ 4 & 1 \end{smallmatrix}\right)$ generate the image of $\Gamma_0(4)$ in $\mathrm{PSL}(2,\mathbb{R})$, i.e., that any element $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(4)$ is, up to sign, a word in T and $\widetilde{T}$. Now a is odd, so $|a| \neq 2|b|$. If $|a| < 2|b|$, then either $b+a$ or $b-a$ is smaller than b in absolute value, so replacing γ by $\gamma \cdot T^{\pm 1}$ decreases a^2+b^2. If $|a| > 2|b| \neq 0$, then either $a+4b$ or $a-4b$ is smaller than a in absolute value, so replacing γ by $\gamma \cdot \widetilde{T}^{\pm 1}$ decreases a^2+b^2. Thus we can keep multiplying γ on the right by powers of T and $\widetilde{T}$ until $b = 0$, at which point $\pm\gamma$ is a power of $\widetilde{T}$.

Now, by the principle "a finite number of q-coefficients suffice" formulated at the end of Section 1, the mere fact that $\theta(z)$ is a modular form is already enough to let one prove non-trivial identities. (We enunciated the principle only in the case of forms of integral weight, but even without knowing the details of the theory it is clear that it then also applies to half-integral weight, since a space of modular forms of half-integral weight can be mapped injectively into a space of modular forms of the next higher integral weight by multiplying by $\theta(z)$.) And indeed, with almost no effort we obtain proofs of two of the most famous results of number theory of the 17th and 18th centuries, the theorems of Fermat and Lagrange about sums of squares.

♠ Sums of Two and Four Squares

Let $r_2(n) = \#\{(a,b) \in \mathbb{Z}^2 \mid a^2+b^2 = n\}$ be the number of representations of an integer $n \geq 0$ as a sum of two squares. Since $\theta(z)^2 = \left(\sum_{a\in\mathbb{Z}} q^{a^2}\right)\left(\sum_{b\in\mathbb{Z}} q^{b^2}\right)$, we see that $r_2(n)$ is simply the coefficient of q^n in $\theta(z)^2$. From Proposition 9 and the just-proved fact that $\Gamma_0(4)$ is generated by $-\mathrm{Id}_2$, T and $\widetilde{T}$, we find that the function $\theta(z)^2$ is a "modular form of weight 1 and character χ_{-4} on $\Gamma_0(4)$" in the sense explained in the paragraph preceding equation (15), where χ_{-4} is the Dirichlet character modulo 4 defined by (15).

Since the Eisenstein series $\mathbb{G}_{1,\chi_{-4}}$ in (16) is also such a modular form, and since the space of all such forms has dimension at most 1 by Proposition 3 (because $\Gamma_0(4)$ has index 6 in $\mathrm{SL}(2,\mathbb{Z})$ and hence volume 2π), these two functions must be proportional. The proportionality factor is obviously 4, and we obtain:

Proposition 10. *Let n be a positive integer. Then the number of representations of n as a sum of two squares is 4 times the sum of $(-1)^{(d-1)/2}$, where d runs over the positive odd divisors of n.*

Corollary (Theorem of Fermat). *Every prime number $p \equiv 1 \pmod 4$ is a sum of two squares.*

Proof of Corollary. We have $r_2(p) = 4\left(1 + (-1)^{(p-1)/4}\right) = 8 \neq 0$.

The same reasoning applies to other powers of θ. In particular, the number $r_4(n)$ of representations of an integer n as a sum of four squares is the coefficient of q^n in the modular form $\theta(z)^4$ of weight 2 on $\Gamma_0(4)$, and the space of all such modular forms is at most two-dimensional by Proposition 3. To find a basis for it, we use the functions $\mathbb{G}_2(z)$ and $\mathbb{G}_2^*(z)$ defined in equations (17) and (21). We showed in §2.3 that the latter function transforms with respect to $\mathrm{SL}(2,\mathbb{Z})$ like a modular form of weight 2, and it follows easily that the three functions $\mathbb{G}_2^*(z)$, $\mathbb{G}_2^*(2z)$ and $\mathbb{G}_2^*(4z)$ transform like modular forms of weight 2 on $\Gamma_0(4)$ (exercise!). Of course these three functions are not holomorphic, but since $\mathbb{G}_2^*(z)$ differs from the holomorphic function $\mathbb{G}_2(z)$ by $1/8\pi y$, we see that the linear combinations $\mathbb{G}_2^*(z)-2\mathbb{G}_2^*(2z) = \mathbb{G}_2(z)-2\mathbb{G}_2(2z)$ and $\mathbb{G}_2^*(2z) - 2\mathbb{G}_2^*(4z) = \mathbb{G}_2(2z) - 2\mathbb{G}_2(4z)$ are holomorphic, and since they are also linearly independent, they provide the desired basis for $M_2(\Gamma_0(4))$. Looking at the first two Fourier coefficients of $\theta(z)^4 = 1 + 8q + \cdots$, we find that $\theta(z)^4$ equals $8\,(\mathbb{G}_2(z)-2\mathbb{G}_2(2z))+16\,(\mathbb{G}_2(2z)-2\mathbb{G}_2(4z))$. Now comparing coefficients of q^n gives:

Proposition 11. *Let n be a positive integer. Then the number of representations of n as a sum of four squares is 8 times the sum of the positive divisors of n which are not multiples of 4.*

Corollary (Theorem of Lagrange). *Every positive integer is a sum of four squares.* ♡

For another simple application of the q-expansion principle, we introduce two variants $\theta_M(z)$ and $\theta_F(z)$ ("M" and "F" for "male" and "female" or "minus sign" and "fermionic") of the function $\theta(z)$ by inserting signs or by shifting the indices by 1/2 in its definition:

$$\begin{aligned}\theta_M(z) &= \sum_{n\in\mathbb{Z}}(-1)^n\, q^{n^2} &&= 1 - 2q + 2q^4 - 2q^9 + \cdots\,,\\ \theta_F(z) &= \sum_{n\in\mathbb{Z}+1/2} q^{n^2} &&= 2q^{1/4} + 2q^{9/4} + 2q^{25/4} + \cdots\,.\end{aligned}$$

These are again modular forms of weight 1/2 on $\varGamma_0(4)$. With a little experimentation, we discover the identity

$$\theta(z)^4 \;=\; \theta_M(z)^4 \,+\, \theta_F(z)^4 \tag{31}$$

due to Jacobi, and by the q-expansion principle all we have to do to prove it is to verify the equality of a finite number of coefficients (here just one). In this particular example, though, there is also an easy combinatorial proof, left as an exercise to the reader.

The three theta series θ, θ_M and θ_F, in a slightly different guise and slightly different notation, play a role in many contexts, so we say a little more about them. As well as the subgroup $\varGamma_0(N)$ of $\varGamma_1$, one also has the *principal congruence subgroup* $\varGamma(N) = \{\gamma \in \varGamma_1 \mid \gamma \equiv \mathrm{Id}_2 \pmod{N}\}$ for every integer $N \in \mathbb{N}$, which is more basic than $\varGamma_0(N)$ because it is a normal subgroup (the kernel of the map $\varGamma_1 \to \mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$ given by reduction modulo N). Exceptionally, the group $\varGamma_0(4)$ is isomorphic to $\varGamma(2)$, simply by conjugation by $\left(\begin{smallmatrix}2&0\\0&1\end{smallmatrix}\right) \in \mathrm{GL}(2,\mathbb{R})$, so that there is a bijection between modular forms on $\varGamma_0(4)$ of any weight and modular forms on $\varGamma(2)$ of the same weight given by $f(z) \to f(z/2)$. In particular, our three theta functions correspond to three new theta functions

$$\theta_3(z) \;=\; \theta(z/2)\,, \qquad \theta_4(z) \;=\; \theta_M(z/2)\,, \qquad \theta_2(z) \;=\; \theta_F(z/2) \tag{32}$$

on $\varGamma(2)$, and the relation (31) becomes $\theta_2^4 + \theta_4^4 = \theta_3^4$. (Here the index “1” is missing because the fourth member of the quartet, $\theta_1(z) = \sum(-1)^n q^{(n+1/2)^2/2}$ is identically zero, as one sees by sending n to $-n-1$. It may look odd that one keeps a whole notation for the zero function. But in fact the functions $\theta_i(z)$ for $1 \le i \le 4$ are just the “Thetanullwerte” or “theta zero-values” of the two-variable series $\theta_i(z,u) = \sum \varepsilon_n\, q^{n^2/2}\, e^{2\pi i n u}$, where the sum is over $\mathbb{Z}$ or $\mathbb{Z} + \frac{1}{2}$ and ε_n is either 1 or $(-1)^n$, none of which vanishes identically. The functions $\theta_i(z,u)$ play a basic role in the theory of elliptic functions and are also the simplest example of *Jacobi forms*, a theory which is related to many of the themes treated in these notes and is also important in connection with Siegel modular forms of degree 2 as discussed in Part III of this book.) The quotient group $\varGamma_1/\varGamma(2) \cong \mathrm{SL}(2,\mathbb{Z}/2\mathbb{Z})$, which has order 6 and is isomorphic to the symmetric group $\mathfrak{S}_3$ on three symbols, acts as the latter on the modular forms $\theta_i(z)^8$, while the fourth powers transform by

$$\varTheta(z) \;:=\; \begin{pmatrix} \theta_2(z)^4 \\ -\,\theta_3(z)^4 \\ \theta_4(z)^4 \end{pmatrix} \quad\Rightarrow\quad \varTheta(z+1) \;=\; -\begin{pmatrix} 1&0&0\\0&0&1\\0&1&0 \end{pmatrix} \varTheta(z),$$

$$z^{-2}\varTheta\left(-\frac{1}{z}\right) \;=\; -\begin{pmatrix} 0&0&1\\0&1&0\\1&0&0 \end{pmatrix} \varTheta(z)\,.$$

This illustrates the general principle that a modular form on a subgroup of finite index of the modular group $\varGamma_1$ can also be seen as a component of a vector-valued modular form on $\varGamma_1$ itself. The full ring $M_*(\varGamma(2))$ of modular forms on $\varGamma(2)$ is generated by the three components of $\varTheta(z)$ (or by any two of them, since their sum is zero), while the subring $M_*(\varGamma(2))^{\mathfrak{S}_3} = M_*(\varGamma_1)$ is generated by the modular forms $\theta_2(z)^8 + \theta_3(z)^8 + \theta_4(z)^8$ and $\bigl(\theta_2(z)^4 + \theta_3(z)^4\bigr)\bigl(\theta_3(z)^4 + \theta_4(z)^4\bigr)\bigl(\theta_4(z)^4 - \theta_2(z)^4\bigr)$ of weights 4 and 6 (which are then equal to $2E_4(z)$ and $2E_6(z)$, respectively). Finally, we see that $\frac{1}{256}\theta_2(z)^8\theta_3(z)^8\theta_4(z)^8$ is a cusp form of weight 12 on $\varGamma_1$ and is, in fact, equal to $\Delta(z)$.

This last identity has an interesting consequence. Since $\Delta(z)$ is non-zero for all $z \in \mathfrak{H}$, it follows that each of the three theta-functions $\theta_i(z)$ has the same property. (One can also see this by noting that the "visible" zero of $\theta_2(z)$ at infinity accounts for all the zeros allowed by the formula discussed in §1.3, so that this function has no zeros at finite points, and then the same holds for $\theta_3(z)$ and $\theta_4(z)$ because they are related to θ_2 by a modular transformation.) This suggests that these three functions, or equivalently their $\varGamma_0(4)$-versions θ, θ_M and θ_F, might have a product expansion similar to that of the function $\Delta(z)$, and indeed this is the case: we have the three identities

$$\theta(z) \;=\; \frac{\eta(2z)^5}{\eta(z)^2\eta(4z)^2}\,, \qquad \theta_M(z) \;=\; \frac{\eta(z)^2}{\eta(2z)}\,, \qquad \theta_F(z) \;=\; 2\,\frac{\eta(4z)^2}{\eta(2z)}\,, \tag{33}$$

where $\eta(z)$ is the "Dedekind eta-function"

$$\eta(z) \;=\; \Delta(z)^{1/24} \;=\; q^{1/24}\prod_{n=1}^{\infty}\bigl(1-q^n\bigr)\,. \tag{34}$$

The proof of (33) is immediate by the usual q-expansion principle: we multiply the identities out (writing, e.g., the first as $\theta(z)\eta(z)^2\eta(4z)^2 = \eta(2z)^5$) and then verify the equality of enough coefficients to account for all possible zeros of a modular form of the corresponding weight. More efficiently, we can use our knowledge of the transformation behavior of $\Delta(z)$ and hence of $\eta(z)$ under $\varGamma_1$ to see that the quotients on the right in (33) are finite at every cusp and hence, since they also have no poles in the upper half-plane, are holomorphic modular forms of weight $1/2$, after which the equality with the theta-functions on the left follows directly.

More generally, one can ask when a quotient of products of eta-functions is a holomorphic modular form. Since $\eta(z)$ is non-zero in $\mathfrak{H}$, such a quotient never has finite zeros, and the only issue is whether the numerator vanishes to at least the same order as the denominator at each cusp. Based on extensive numerical calculations, I formulated a general conjecture saying that there are essentially only finitely many such products of any given weight, and a second explicit conjecture giving the complete list for weight $1/2$ (i.e., when the number of η's in the numerator is one bigger than in the denominator). Both conjectures were proved by Gerd Mersmann in a brilliant Master's thesis.

For the weight 1/2 result, the meaning of "essentially" is that the product should be primitive, i.e., it should have the form $\prod \eta(n_i z)^{a_i}$ where the n_i are positive integers with no common factor. (Otherwise one would obtain infinitely many examples by rescaling, e.g., one would have both $\theta_M(z) = \eta(z)^2/\eta(2z)$ and $\theta_M(2z) = \eta(2z)^2/\eta(4z)$ on the list.) The classification is then as follows:

Theorem (Mersmann). *There are precisely* 14 *primitive eta-products which are holomorphic modular forms of weight* 1/2 *:*

$$\eta(z)\,,\quad \frac{\eta(z)^2}{\eta(2z)}\,,\quad \frac{\eta(2z)^2}{\eta(z)}\,,\quad \frac{\eta(z)\,\eta(4z)}{\eta(2z)}\,,\quad \frac{\eta(2z)^3}{\eta(z)\,\eta(4z)}\,,\quad \frac{\eta(2z)^5}{\eta(z)^2\eta(4z)^2}\,,$$
$$\frac{\eta(z)^2\eta(6z)}{\eta(2z)\,\eta(3z)}\,,\quad \frac{\eta(2z)^2\eta(3z)}{\eta(z)\,\eta(6z)}\,,\quad \frac{\eta(2z)\,\eta(3z)^2}{\eta(z)\,\eta(6z)}\,,\quad \frac{\eta(z)\,\eta(6z)^2}{\eta(2z)\,\eta(3z)}\,,$$
$$\frac{\eta(z)\,\eta(4z)\,\eta(6z)^2}{\eta(2z)\,\eta(3z)\,\eta(12z)}\,,\quad \frac{\eta(2z)^2\eta(3z)\,\eta(12z)}{\eta(z)\,\eta(4z)\,\eta(6z)}\,,\quad \frac{\eta(2z)^5\eta(3z)\,\eta(12z)}{\eta(z)^2\eta(4z)^2\eta(6z)^2}\,,$$
$$\frac{\eta(z)\,\eta(4z)\,\eta(6z)^5}{\eta(2z)^2\eta(3z)^2\eta(12z)^2}\,.$$

Finally, we mention that $\eta(z)$ itself has the theta-series representation

$$\eta(z) \;=\; \sum_{n=1}^{\infty} \chi_{12}(n)\, q^{n^2/24} \;=\; q^{1/24} \,-\, q^{25/24} \,-\, q^{49/24} \,+\, q^{121/24} \,+\, \cdots$$

where $\chi_{12}(12m \pm 1) = 1$, $\chi_{12}(12m \pm 5) = -1$, and $\chi_{12}(n) = 0$ if n is divisible by 2 or 3. This identity was discovered numerically by Euler (in the simpler-looking but less enlightening version $\prod_{n=1}^{\infty}(1-q^n) = \sum_{n=1}^{\infty}(-1)^n\, q^{(3n^2+n)/2}$) and proved by him only after several years of effort. From a modern point of view, his theorem is no longer surprising because one now knows the following beautiful general result, proved by J-P. Serre and H. Stark in 1976:

Theorem (Serre–Stark). *Every modular form of weight 1/2 is a linear combination of unary theta series.*

Explicitly, this means that every modular form of weight 1/2 with respect to any subgroup of finite index of $\mathrm{SL}(2,\mathbb{Z})$ is a linear combination of sums of the form $\sum_{n\in\mathbb{Z}} q^{a(n+c)^2}$ with $a \in \mathbb{Q}_{>0}$ and $c \in \mathbb{Q}$. Euler's formula for $\eta(z)$ is a typical case of this, and of course each of the other products given in Mersmann's theorem must also have a representation as a theta series. For instance, the last function on the list, $\eta(z)\eta(4z)\eta(6z)^5/\eta(2z)^2\eta(3z)^2\eta(12z)^2$, has the expansion $\sum_{n>0,\,(n,6)=1} \chi_8(n) q^{n^2/24}$, where $\chi_8(n)$ equals $+1$ for $n \equiv \pm 1 \pmod 8$ and -1 for $n \equiv \pm 3 \pmod 8$.

We end this subsection by mentioning one more application of the Jacobi theta series.

♠ The Kac–Wakimoto Conjecture

For any two natural numbers m and n, denote by $\Delta_m(n)$ the number of representations of n as a sum of m triangular numbers (numbers of the form $a(a-1)/2$ with a integral). Since $8a(a-1)/2+1=(2a-1)^2$, this can also be written as the number $r_m^{\mathrm{odd}}(8n+m)$ of representations of $8n+m$ as a sum of m odd squares. As part of an investigation in the theory of affine superalgebras, Kac and Wakimoto were led to conjecture the formula

$$\Delta_{4s^2}(n) \;=\; \sum_{\substack{r_1,\,a_1,\,\dots,\,r_s,\,a_s\,\in\,\mathbb{N}_{\mathrm{odd}}\\ r_1a_1+\cdots+r_sa_s\,=\,2n+s^2}} P_s(a_1,\dots,a_s) \tag{35}$$

for m of the form $4s^2$ (and a similar formula for m of the form $4s(s+1)$), where $\mathbb{N}_{\mathrm{odd}}=\{1,3,5,\dots\}$ and P_s is the polynomial

$$P_s(a_1,\dots,a_s) \;=\; \frac{\prod_i a_i\,\cdot\,\prod_{i<j}\bigl(a_i^2-a_j^2\bigr)^2}{4^{s(s-1)}\,s!\,\prod_{j=1}^{2s-1}j!}\,.$$

Two proofs of this were subsequently given, one by S. Milne using elliptic functions and one by myself using modular forms. Milne's proof is very ingenious, with a number of other interesting identities appearing along the way, but is quite involved. The modular proof is much simpler. One first notes that, P_s being a homogeneous polynomial of degree $2s^2-s$ and odd in each argument, the right-hand side of (35) is the coefficient of q^{2n+s^2} in a function $F(z)$ which is a linear combination of products $g_{h_1}(z)\cdots g_{h_s}(z)$ with $h_1+\cdots+h_s=s^2$, where $g_h(z)=\sum_{r,\,a\in\mathbb{N}_{\mathrm{odd}}}a^{2h-1}q^{ra}$ $(h\ge 1)$. Since g_h is a modular form (Eisenstein series) of weight $2h$ on $\Gamma_0(4)$, this function F is a modular form of weight $2s^2$ on the same group. Moreover, its Fourier expansion belongs to $q^{s^2}\mathbb{Q}[[q^2]]$ (because $P_s(a_1,\dots,a_s)$ vanishes if any two a_i are equal, and the smallest value of $r_1a_1+\cdots+r_sa_s$ with all r_i and a_i in $\mathbb{N}_{\mathrm{odd}}$ and all a_i distinct is $1+3+\cdots+2s-1=s^2$), and from the formula given in §1 for the number of zeros of a modular form we find that this property characterizes $F(z)$ uniquely in $M_{2s^2}(\Gamma_0(4))$ up to a scalar factor. But $\theta_F(z)^{4s^2}$ has the same property, so the two functions must be proportional. This proves (35) up to a scalar factor, easily determined by setting $n=0$. ♡

3.2 Theta Series in Many Variables

We now consider quadratic forms in an arbitrary number m of variables. Let $Q:\mathbb{Z}^m\to\mathbb{Z}$ be a positive definite quadratic form which takes integral values on $\mathbb{Z}^m$. We associate to Q the theta series

$$\Theta_Q(z) \;=\; \sum_{x_1,\dots,x_m\in\mathbb{Z}} q^{Q(x_1,\dots,x_m)} \;=\; \sum_{n=0}^{\infty} R_Q(n)\,q^n\,, \tag{36}$$

where of course $q = e^{2\pi i z}$ as usual and $R_Q(n) \in \mathbb{Z}_{\geq 0}$ denotes the number of representations of n by Q, i.e., the number of vectors $x \in \mathbb{Z}^m$ with $Q(x) = n$. The basic statement is that Θ_Q is always a modular form of weight $m/2$. In the case of even m we can be more precise about the modular transformation behavior, since then we are in the realm of modular forms of integral weight where we have given complete definitions of what modularity means. The quadratic form $Q(x)$ is a linear combination of products $x_i x_j$ with $1 \leq i,\, j \leq m$. Since $x_i x_j = x_j x_i$, we can write $Q(x)$ uniquely as

$$Q(x) \;=\; \frac{1}{2}\,x^t A x \;=\; \frac{1}{2}\sum_{i,j=1}^{m} a_{ij} x_i x_j\,, \tag{37}$$

where $A = (a_{ij})_{1 \leq i,j \leq m}$ is a symmetric $m \times m$ matrix and the factor $1/2$ has been inserted to avoid counting each term twice. The integrality of Q on $\mathbb{Z}^m$ is then equivalent to the statement that the symmetric matrix A has integral elements and that its diagonal elements a_{ii} are even. Such an A is called an *even integral matrix.* Since we want $Q(x) > 0$ for $x \neq 0$, the matrix A must be positive definite. This implies that $\det A > 0$. Hence A is non-singular and A^{-1} exists and belongs to $M_m(\mathbb{Q})$. The *level* of Q is then defined as the smallest positive integer $N = N_Q$ such that NA^{-1} is again an even integral matrix. We also have the *discriminant* $\Delta = \Delta_Q$ of A, defined as $(-1)^m \det A$. It is always congruent to 0 or 1 modulo 4, so there is an associated character (Kronecker symbol) χ_Δ, which is the unique Dirichlet character modulo N satisyfing $\chi_\Delta(p) = \left(\frac{\Delta}{p}\right)$ (Legendre symbol) for any odd prime $p \nmid N$. (The character χ_Δ in the special cases $\Delta = -4$, 12 and 8 already occurred in §2.2 (eq. (15)) and §3.1.) The precise description of the modular behavior of Θ_Q for $m \in 2\mathbb{Z}$ is then:

Theorem (Hecke, Schoenberg). *Let $Q : \mathbb{Z}^{2k} \to \mathbb{Z}$ be a positive definite integer-valued form in $2k$ variables of level N and discriminant Δ. Then Θ_Q is a modular form on $\Gamma_0(N)$ of weight k and character χ_Δ, i.e., we have $\Theta_Q\bigl(\frac{az+b}{cz+d}\bigr) = \chi_\Delta(a)\,(cz+d)^k\,\Theta_Q(z)$ for all $z \in \mathfrak{H}$ and $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)$.*

The proof, as in the unary case, relies essentially on the Poisson summation formula, which gives the identity $\Theta_Q(-1/Nz) = N^{k/2}(z/i)^k\,\Theta_{Q^*}(z)$, where $Q^*(x)$ is the quadratic form associated to NA^{-1}, but finding the precise modular behavior requires quite a lot of work. One can also in principle reduce the higher rank case to the one-variable case by using the fact that every quadratic form is diagonalizable over $\mathbb{Q}$, so that the sum in (36) can be broken up into finitely many sub-sums over sublattices or translated sublattices of $\mathbb{Z}^m$ on which $Q(x_1, \ldots, x_m)$ can be written as a linear combination of m squares.

There is another language for quadratic forms which is often more convenient, the language of lattices. From this point of view, a quadratic form is no longer a homogeneous quadratic polynomial in m variables, but a function Q

from a free $\mathbb{Z}$-module Λ of rank m to $\mathbb{Z}$ such that the associated scalar product $(x,y) = Q(x+y) - Q(x) - Q(y)$ $(x,\, y \in \Lambda)$ is bilinear. Of course we can always choose a $\mathbb{Z}$-basis of Λ, in which case Λ is identified with $\mathbb{Z}^m$ and Q is described in terms of a symmetric matrix A as in (37), the scalar product being given by $(x,y) = x^t Ay$, but often the basis-free language is more convenient. In terms of the scalar product, we have a length function $\|x\|^2 = (x,x)$ (actually this is the square of the length, but one often says simply "length" for convenience) and $Q(x) = \frac{1}{2}\|x\|^2$, so that the integer-valued case we are considering corresponds to lattices in which all vectors have even length. One often chooses the lattice Λ inside the euclidean space $\mathbb{R}^m$ with its standard length function $(x,x) = \|x\|^2 = x_1^2 + \cdots + x_m^2$; in this case the square root of $\det A$ is equal to the volume of the quotient $\mathbb{R}^m/\Lambda$, i.e., to the volume of a fundamental domain for the action by translation of the lattice Λ on $\mathbb{R}^m$. In the case when this volume is 1, i.e., when $\Lambda \in \mathbb{R}^m$ has the same covolume as $\mathbb{Z}^m$, the lattice is called *unimodular*. Let us look at this case in more detail.

♠ Invariants of Even Unimodular Lattices

If the matrix A in (37) is even and unimodular, then the above theorem tells us that the theta series Θ_Q associated to Q is a modular form on the full modular group. This has many consequences.

Proposition 12. *Let $Q : \mathbb{Z}^m \to \mathbb{Z}$ be a positive definite even unimodular quadratic form in m variables. Then*

(i) the rank m is divisible by 8, and
(ii) the number of representations of $n \in \mathbb{N}$ by Q is given for large n by the formula

$$R_Q(n) \;=\; -\frac{2k}{B_k}\,\sigma_{k-1}(n) \;+\; O\bigl(n^{k/2}\bigr) \qquad (n \to \infty)\,, \tag{38}$$

where $m = 2k$ and B_k denotes the kth Bernoulli number.

Proof. For the first part it is enough to show that m cannot be an odd multiple of 4, since if m is either odd or twice an odd number then $4m$ or $2m$ is an odd multiple of 4 and we can apply this special case to the quadratic form $Q \oplus Q \oplus Q \oplus Q$ or $Q \oplus Q$, respectively. So we can assume that $m = 2k$ with k even and must show that k is divisible by 4 and that (38) holds. By the theorem above, the theta series Θ_Q is a modular form of weight k on the full modular group $\Gamma_1 = \mathrm{SL}(2,\mathbb{Z})$ (necessarily with trivial character, since there are no non-trivial Dirichlet characters modulo 1). By the results of Section 2, this modular form is a linear combination of $\mathbb{G}_k(z)$ and a cusp form of weight k, and from the Fourier expansion (13) we see that the coefficient of $\mathbb{G}_k$ in this decomposition equals $-2k/B_k$, since the constant term $R_Q(0)$ of Θ_Q equals 1. (The only vector of length 0 is the zero vector.) Now Proposition 8 implies the asymptotic formula (38), and the fact that k must be divisible by 4 also follows because if $k \equiv 2 \pmod 4$ then B_k is positive and therefore the right-hand side of (38) tends to $-\infty$ as $k \to \infty$, contradicting $R_Q(n) \geq 0$.

The first statement of Proposition 12 is purely algebraic, and purely algebraic proofs are known, but they are not as simple or as elegant as the modular proof just given. No non-modular proof of the asymptotic formula (38) is known.

Before continuing with the theory, we look at some examples, starting in rank 8. Define the lattice $\Lambda_8 \subset \mathbb{R}^8$ to be the set of vectors belonging to either $\mathbb{Z}^8$ or $(\mathbb{Z}+\frac{1}{2})^8$ for which the sum of the coordinates is even. This is unimodular because the lattice $\mathbb{Z}^8 \cup (\mathbb{Z}+\frac{1}{2})^8$ contains both it and $\mathbb{Z}^8$ with the same index 2, and is even because $x_i^2 \equiv x_i \pmod 2$ for $x_i \in \mathbb{Z}$ and $x_i^2 \equiv \frac{1}{4} \pmod 2$ for $x_i \in \mathbb{Z}+\frac{1}{2}$. The lattice Λ_8 is sometimes denoted E_8 because, if we choose the $\mathbb{Z}$-basis $u_i = e_i - e_{i+1}$ $(1 \leq i \leq 6)$, $u_7 = e_6 + e_7$, $u_8 = -\frac{1}{2}(e_1 + \cdots + e_8)$ of Λ_8, then every u_i has length 2 and (u_i, u_j) for $i \neq j$ equals -1 or 0 according whether the ith and jth vertices (in a standard numbering) of the "E_8" Dynkin diagram in the theory of Lie algebras are adjacent or not. The theta series of Λ_8 is a modular form of weight 4 on $\mathrm{SL}(2,\mathbb{Z})$ whose Fourier expansion begins with 1, so it is necessarily equal to $E_4(z)$, and we get "for free" the information that for every integer $n \geq 1$ there are exactly $240\,\sigma_3(n)$ vectors x in the E_8 lattice with $(x,x) = 2n$.

From the uniqueness of the modular form $E_4 \in M_4(\Gamma_1)$ we in fact get that $r_Q(n) = 240\sigma_3(n)$ for any even unimodular quadratic form or lattice of rank 8, but here this is not so interesting because the known classification in this rank says that Λ_8 is, in fact, the only such lattice up to isomorphism. However, in rank 16 one knows that there are two non-equivalent lattices: the direct sum $\Lambda_8 \oplus \Lambda_8$ and a second lattice Λ_{16} which is not decomposable. Since the theta series of both lattices are modular forms of weight 8 on the full modular group with Fourier expansions beginning with 1, they are both equal to the Eisenstein series $E_8(z)$, so we have $r_{\Lambda_8 \oplus \Lambda_8}(n) = r_{\Lambda_{16}}(n) = 480\,\sigma_7(n)$ for all $n \geq 1$, even though the two lattices in question are distinct. (Their distinctness, and a great deal of further information about the relative positions of vectors of various lengths in these or in any other lattices, can be obtained by using the theory of Jacobi forms which was mentioned briefly in §3.1 rather than just the theory of modular forms.)

In rank 24, things become more interesting, because now $\dim M_{12}(\Gamma_1) = 2$ and we no longer have uniqueness. The even unimodular lattices of this rank were classified completely by Niemeyer in 1973. There are exactly 24 of them up to isomorphism. Some of them have the same theta series and hence the same number of vectors of any given length (an obvious such pair of lattices being $\Lambda_8 \oplus \Lambda_8 \oplus \Lambda_8$ and $\Lambda_8 \oplus \Lambda_{16}$), but not all of them do. In particular, exactly one of the 24 lattices has the property that it has no vectors of length 2. This is the famous Leech lattice (famous among other reasons because it has a huge group of automorphisms, closely related to the monster group and other sporadic simple groups). Its theta series is the unique modular form of weight 12 on Γ_1 with Fourier expansion starting $1 + 0q + \cdots$, so it must equal $E_{12}(z) - \frac{21736}{691}\Delta(z)$, i.e., the number $r_{\text{Leech}}(n)$ of vectors of length $2n$ in the Leech lattice equals $\frac{21736}{691}\bigl(\sigma_{11}(n) - \tau(n)\bigr)$ for every positive integer n. This gives another proof and an interpretation of Ramanujan's congruence (28).

In rank 32, things become even more interesting: here the complete classification is not known, and we know that we cannot expect it very soon, because there are more than 80 million isomorphism classes! This, too, is a consequence of the theory of modular forms, but of a much more sophisticated part than we are presenting here. Specifically, there is a fundamental theorem of Siegel saying that the average value of the theta series associated to the quadratic forms in a single genus (we omit the definition) is always an Eisenstein series. Specialized to the particular case of even unimodular forms of rank $m = 2k \equiv 0 \pmod 8$, which form a single genus, this theorem says that there are only finitely many such forms up to equivalence for each k and that, if we number them $Q_1, \ldots, Q_I$, then we have the relation

$$\sum_{i=1}^{I} \frac{1}{w_i}\,\Theta_{Q_i}(z) \;=\; \mathsf{m}_k\, E_k(z)\,, \tag{39}$$

where w_i is the number of automorphisms of the form Q_i (i.e., the number of matrices $\gamma \in \mathrm{SL}(m,\mathbb{Z})$ such that $Q_i(\gamma x) = Q_i(x)$ for all $x \in \mathbb{Z}^m$) and m_k is the positive rational number given by the formula

$$\mathsf{m}_k \;=\; \frac{B_k}{2k}\,\frac{B_2}{4}\,\frac{B_4}{8}\,\cdots\,\frac{B_{2k-2}}{4k-4}\,,$$

where B_i denotes the ith Bernoulli number. In particular, by comparing the constant terms on the left- and right-hand sides of (39), we see that $\sum_{i=1}^{I} 1/w_i = \mathsf{m}_k$, the *Minkowski-Siegel mass formula.* The numbers $\mathsf{m}_4 \approx 1.44 \times 10^{-9}$, $\mathsf{m}_8 \approx 2.49 \times 10^{-18}$ and $\mathsf{m}_{12} \approx 7{,}94 \times 10^{-15}$ are small, but $\mathsf{m}_{16} \approx 4{,}03 \times 10^{7}$ (the next two values are $\mathsf{m}_{20} \approx 4.39 \times 10^{51}$ and $\mathsf{m}_{24} \approx 1.53 \times 10^{121}$), and since $w_i \geq 2$ for every i (one has at the very least the automorphisms $\pm\,\mathrm{Id}_m$), this shows that $I > 80000000$ for $m = 32$ as asserted.

A further consequence of the fact that $\Theta_Q \in M_k(\Gamma_1)$ for Q even and unimodular of rank $m = 2k$ is that the minimal value of $Q(x)$ for non-zero $x \in \Lambda$ is bounded by $r = \dim M_k(\Gamma_1) = [k/12] + 1$. The lattice L is called *extremal* if this bound is attained. The three lattices of rank 8 and 16 are extremal for trivial reasons. (Here $r = 1$.) For $m = 24$ we have $r = 2$ and the only extremal lattice is the Leech lattice. Extremal unimodular lattices are also known to exist for $m = 32$, 40, 48, 56, 64 and 80, while the case $m = 72$ is open. Surprisingly, however, there are no examples of large rank:

Theorem (Mallows–Odlyzko–Sloane). *There are only finitely many non-isomorphic extremal even unimodular lattices.*

We sketch the proof, which, not surprisingly, is completely modular. Since there are only finitely many non-isomorphic even unimodular lattices of any given rank, the theorem is equivalent to saying that there is an absolute bound on the value of the rank m for extremal lattices. For simplicity, let us suppose that $m = 24n$. (The cases $m = 24n + 8$ and $m = 24n + 16$ are similar.) The theta series of any extremal unimodular lattice of this rank must be equal to the unique modular form $f_n \in M_{12n}(\mathrm{SL}(2, \mathbb{Z}))$ whose q-development has the form $1 + \mathrm{O}(q^{n+1})$. By an elementary argument which we omit but which the reader may want to look for, we find that this q-development has the form

$$f_n(z) \;=\; 1 \;+\; n\, a_n\, q^{n+1} \;+\; \left(\frac{nb_n}{2} \;-\; 24\, n\, (n+31)\, a_n\right) q^{n+2} \;+\; \cdots$$

where a_n and b_n are the coefficients of $\Delta(z)^n$ in the modular functions $j(z)$ and $j(z)^2$, respectively, when these are expressed (locally, for small q) as Laurent series in the modular form $\Delta(z) = q - 24q^2 + 252q^3 - \cdots$. It is not hard to show that a_n has the asymptotic behavior $a_n \sim An^{-3/2}C^n$ for some constants $A = 225153.793389\cdots$ and $C = 1/\Delta(z_0) = 69.1164201716\cdots$, where $z_0 = 0.52352170017992\cdots i$ is the unique zero on the imaginary axis of the function $E_2(z)$ defined in (17) (this is because $E_2(z)$ is the logarithmic derivative of $\Delta(z)$), while b_n has a similar expansion but with A replaced by $2\lambda A$ with $\lambda = j(z_0) - 720 = 163067.793145\cdots$. It follows that the coefficient $\frac{1}{2}nb_n - 24n(n+31)a_n$ of q^{n+2} in f_n is negative for n larger than roughly 6800, corresponding to $m \approx 163000$, and that therefore extremal lattices of rank larger than this cannot exist. ♡

♠ Drums Whose Shape One Cannot Hear

Marc Kac asked a famous question, “Can one hear the shape of a drum?” Expressed more mathematically, this means: can there be two riemannian manifolds (in the case of real “drums” these would presumably be two-dimensional manifolds with boundary) which are not isometric but have the same spectra of eigenvalues of their Laplace operators? The first example of such a pair of manifolds to be found was given by Milnor, and involved 16-dimensional closed “drums.” More drum-like examples consisting of domains in $\mathbb{R}^2$ with polygonal boundary are now also known, but they are difficult to construct, whereas Milnor's example is very easy. It goes as follows. As we already mentioned, there are two non-isomorphic even unimodular lattices $\Lambda_1 = \Gamma_8 \oplus \Gamma_8$ and $\Lambda_2 = \Gamma_{16}$ in dimension 16. The fact that they are non-isomorphic means that the two Riemannian manifolds $M_1 = \mathbb{R}^{16}/\Lambda_1$ and $M_2 = \mathbb{R}^{16}/\Lambda_2$, which are topologically both just tori $(S^1)^{16}$, are not isometric to each other. But the spectrum of the Laplace operator on any torus $\mathbb{R}^n/\Lambda$ is just the set of norms $\|\lambda\|^2$ ($\lambda \in \Lambda$), counted with multiplicities, and these spectra agree for M_1 and M_2 because the theta series $\sum_{\lambda\in\Lambda_1} q^{\|\lambda\|^2}$ and $\sum_{\lambda\in\Lambda_2} q^{\|\lambda\|^2}$ coincide. ♡

We should not leave this section without mentioning at least briefly that there is an important generalization of the theta series (36) in which each term $q^{Q(x_1,\ldots,x_m)}$ is weighted by a polynomial $P(x_1,\ldots,x_m)$. If this polynomial is homogeneous of degree d and is *spherical with respect to* Q (this means that $\Delta P = 0$, where Δ is the Laplace operator with respect to a system of coordinates in which $Q(x_1,\ldots,x_m)$ is simply $x_1^2 + \cdots + x_m^2$), then the theta series $\Theta_{Q,P}(z) = \sum_x P(x)q^{Q(x)}$ is a modular form of weight $m/2 + d$ (on the same group and with respect to the same character as in the case $P = 1$), and is a cusp form if d is strictly positive. The possibility of putting non-trivial weights into theta series in this way considerably enlarges their range of applications, both in coding theory and elsewhere.

4 Hecke Eigenforms and L-series

In this section we give a brief sketch of Hecke's fundamental discoveries that the space of modular forms is spanned by modular forms with multiplicative Fourier coefficients and that one can associate to these forms Dirichlet series which have Euler products and functional equations. These facts are at the basis of most of the higher developments of the theory: the relations of modular forms to arithmetic algebraic geometry and to the theory of motives, and the adelic theory of automorphic forms. The last two subsections describe some basic examples of these higher connections.

4.1 Hecke Theory

For each integer $m \geq 1$ there is a linear operator T_m, the mth *Hecke operator*, acting on modular forms of any given weight k. In terms of the description of modular forms as homogeneous functions on lattices which was given in §1.1, the definition of T_m is very simple: it sends a homogeneous function F of degree $-k$ on lattices $\Lambda \subset \mathbb{C}$ to the function $T_m F$ defined (up to a suitable normalizing constant) by $T_m F(\Lambda) = \sum F(\Lambda')$, where the sum runs over all sublattices $\Lambda' \subset \Lambda$ of index m. The sum is finite and obviously still homogeneous in Λ of the same degree $-k$. Translating from the language of lattices to that of functions in the upper half-plane by the usual formula $f(z) = F(\Lambda_z)$, we find that the action of T_m is given by

$$T_m f(z) \;=\; m^{k-1} \sum_{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_1 \backslash \mathcal{M}_m} (cz+d)^{-k}\, f\left(\frac{az+b}{cz+d}\right) \qquad (z \in \mathfrak{H}), \quad (40)$$

where $\mathcal{M}_m$ denotes the set of 2×2 integral matrices of determinant m and where the normalizing constant m^{k-1} has been introduced for later convenience (T_m normalized in this way will send forms with integral Fourier coefficients to forms with integral Fourier coefficients). The sum makes sense because the transformation law (2) of f implies that the summand associated to a matrix $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathcal{M}_m$ is indeed unchanged if M is replaced by γM with $\gamma \in \varGamma_1$, and from (40) one also easily sees that $T_m f$ is holomorphic in $\mathfrak{H}$ and satisfies the same transformation law and growth properties as f, so T_m indeed maps $M_k(\varGamma_1)$ to $M_k(\varGamma_1)$. Finally, to calculate the effect of T_m on Fourier developments, we note that a set of representatives of $\varGamma_1\backslash\mathcal{M}_m$ is given by the upper triangular matrices $\left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right)$ with $ad = m$ and $0 \le b < d$ (this is an easy exercise), so

$$T_m f(z) \;=\; m^{k-1} \sum_{\substack{ad=m \\ a,\,d>0}} \frac{1}{d^k} \sum_{b \pmod{d}} f\left(\frac{az+b}{d}\right) . \tag{41}$$

If $f(z)$ has the Fourier development (3), then a further calculation with (41), again left to the reader, shows that the function $T_m f(z)$ has the Fourier expansion

$$T_m f(z) \;=\; \sum_{\substack{d\mid m \\ d>0}} (m/d)^{k-1} \sum_{\substack{n\ge 0 \\ d\mid n}} a_n\, q^{mn/d^2} \;=\; \sum_{n\ge 0} \biggl(\sum_{\substack{r\mid(m,n) \\ r>0}} r^{k-1}\, a_{mn/r^2} \biggr) q^n \,. \tag{42}$$

An easy but important consequence of this formula is that the operators T_m ($m \in \mathbb{N}$) all commute.

Let us consider some examples. The expansion (42) begins $\sigma_{k-1}(m)a_0 + a_m q + \cdots$, so if f is a cusp form (i.e., $a_0 = 0$), then so is $T_m f$. In particular, since the space $S_{12}(\varGamma_1)$ of cusp forms of weight 12 is 1-dimensional, spanned by $\Delta(z)$, it follows that $T_m\Delta$ is a multiple of Δ for every $m \ge 1$. Since the Fourier expansion of Δ begins $q + \cdots$ and that of $T_m\Delta$ begins $\tau(m)q + \cdots$, the eigenvalue is necessarily $\tau(m)$, so $T_m\Delta = \tau(m)\Delta$ and (42) gives

$$\tau(m)\,\tau(n) \;=\; \sum_{r\mid(m,n)} r^{11}\,\tau\left(\frac{mn}{r^2}\right) \qquad \text{for all } m,\, n \ge 1\,,$$

proving Ramanujan's multiplicativity observations mentioned in §2.4. By the same argument, if $f \in M_k(\varGamma_1)$ is any simultaneous eigenfunction of all of the T_m, with eigenvalues λ_m, then $a_m = \lambda_m a_1$ for all m. We therefore have $a_1 \ne 0$ if f is not identically 0, and if we normalize f by $a_1 = 1$ (such an f is called a *normalized Hecke eigenform*, or *Hecke form* for short) then we have

$$T_m f \;=\; a_m\, f\,, \qquad a_m\, a_n \;=\; \sum_{r\mid(m,n)} r^{k-1}\, a_{mn/r^2} \qquad (m,\, n \ge 1)\,. \tag{43}$$

Examples of this besides $\Delta(z)$ are the unique normalized cusp forms $f(z) = \Delta(z)E_{k-12}(z)$ in the five further weights where $\dim S_k(\varGamma_1) = 1$ (viz. $k = 16$, 18, 20, 22 and 26) and the function $\mathbb{G}_k(z)$ for all $k \ge 4$, for which we have $T_m\mathbb{G}_k = \sigma_{k-1}(m)\mathbb{G}_k$, $\sigma_{k-1}(m)\sigma_{k-1}(n) = \sum_{r\mid(m,n)} r^{k-1}\sigma_{k-1}(mn/r^2)$. (This was the reason for the normalization of $\mathbb{G}_k$ chosen in §2.2.) In fact, a theorem of Hecke asserts that $M_k(\Gamma_1)$ has a basis of normalized simultaneous eigenforms for all k, and that this basis is unique. We omit the proof, though it is not difficult (one introduces a scalar product on the space of cusp forms of weight k, shows that the T_m are self-adjoint with respect to this scalar product, and appeals to a general result of linear algebra saying that commuting self-adjoint operators can always be simultaneously diagonalized), and content ourselves instead with one further example, also due to Hecke. Consider $k = 24$, the first weight where $\dim S_k(\Gamma_1)$ is greater than 1. Here S_k is 2-dimensional, spanned by $\Delta E_4^3 = q + 696q^2 + \cdots$ and $\Delta^2 = q^2 - 48q^3 + \ldots$. Computing the first two Fourier expansions of the images under T_2 of these two functions by (42), we find that $T_2(\Delta E_4^3) = 696\,\Delta E_4^3 + 20736000\,\Delta^2$ and $T_2(\Delta^2) = \Delta E_4^3 + 384\,\Delta^2$. The matrix $\left(\begin{smallmatrix} 696 & 20736000 \\ 1 & 384 \end{smallmatrix}\right)$ has distinct eigenvalues $\lambda_1 = 540 + 12\sqrt{144169}$ and $\lambda_2 = 540 - 12\sqrt{144169}$, so there are precisely two normalized eigenfunctions of T_2 in $S_{24}(\Gamma_1)$, namely the functions $f_1 = \Delta E_4^3 - (156 - 12\sqrt{144169})\Delta^2 = q + \lambda_1 q^2 + \cdots$ and $f_2 = \Delta E_4^3 - (156 + 12\sqrt{144169})\Delta^2 = q + \lambda_2 q^2 + \cdots$, with $T_2 f_i = \lambda_i f_i$ for $i = 1, 2$. The uniqueness of these eigenfunctions and the fact that T_m commutes with T_2 for all $m \ge 1$ then implies that $T_m f_i$ is a multiple of f_i for all $m \ge 1$, so $\mathbb{G}_{24}$, f_1 and f_2 give the desired unique basis of $M_{24}(\Gamma_1)$ consisting of normalized Hecke eigenforms.

Finally, we mention without giving any details that Hecke's theory generalizes to congruence groups of $\mathrm{SL}(2,\mathbb{Z})$ like the group $\Gamma_0(N)$ of matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_1$ with $c \equiv 0 \pmod{N}$, the main differences being that the definition of T_m must be modified somewhat if m and N are not coprime and that the statement about the existence of a unique base of Hecke forms becomes more complicated: the space $M_k(\Gamma_0(N))$ is the direct sum of the space spanned by all functions $f(dz)$ where $f \in M_k(\Gamma_0(N'))$ for some proper divisor N' of N and d divides N/N' (the so-called "old forms") and a space of "new forms" which is again uniquely spanned by normalized eigenforms of all Hecke operators T_m with $(m, N) = 1$. The details can be found in any standard textbook.

4.2 *L*-series of Eigenforms

Let us return to the full modular group. We have seen that $M_k(\Gamma_1)$ contains, and is in fact spanned by, normalized Hecke eigenforms $f = \sum a_m q^m$ satisfying (43). Specializing this equation to the two cases when m and n are coprime and when $m = p^\nu$ and $n = p$ for some prime p gives the two equations (which together are equivalent to (43))

$$a_{mn} = a_m\, a_n \text{ if } (m,n) = 1\,, \quad a_{p^{\nu+1}} = a_p\, a_{p^\nu} - p^{k-1}\, a_{p^{\nu-1}} \quad (p \text{ prime}, \nu \ge 1)\,.$$

The first says that the coefficients a_n are *multiplicative* and hence that the Dirichlet series $L(f,s) = \sum\limits_{n=1}^{\infty} \dfrac{a_n}{n^s}$, called the *Hecke L-series* of f, has an Euler product $L(f,s) = \prod_{p \text{ prime}} \bigl(1 + \frac{a_p}{p^s} + \frac{a_{p^2}}{p^{2s}} + \cdots\bigr)$, and the second tells us that the power series $\sum_{\nu=0}^{\infty} a_{p^\nu} x^\nu$ for p prime equals $1/(1 - a_p x + p^{k-1}x^2)$. Combining these two statements gives Hecke's fundamental Euler product development

$$L(f,s) \;=\; \prod_{p \text{ prime}} \frac{1}{1 \,-\, a_p\, p^{-s} \,+\, p^{k-1-2s}} \tag{44}$$

for the L-series of a normalized Hecke eigenform $f \in M_k(\Gamma_1)$, a simple example being given by

$$L(\mathbb{G}_k, s) \;=\; \prod_p \frac{1}{1 - (p^{k-1}+1)p^{-s} + p^{k-1-2s}} \;=\; \zeta(s)\,\zeta(s-k+1)\,.$$

For eigenforms on $\Gamma_0(N)$ there is a similar result except that the Euler factors for $p|N$ have to be modified suitably.

The L-series have another fundamental property, also discovered by Hecke, which is that they can be analytically continued in s and then satisfy functional equations. We again restrict to $\Gamma = \Gamma_1$ and also, for convenience, to cusp forms, though not any more just to eigenforms. (The method of proof extends to non-cusp forms but is messier there since $L(f,s)$ then has poles, and since M_k is spanned by cusp forms and by $\mathbb{G}_k$, whose L-series is completely known, there is no loss in making the latter restriction.) From the estimate $a_n = \mathrm{O}(n^{k/2})$ proved in §2.4 we know that $L(f,s)$ converges absolutely in the half-plane $\Re(s) > 1 + k/2$. Take s in that half-plane and consider the Euler gamma function

$$\Gamma(s) \;=\; \int_0^\infty t^{s-1}\, e^{-t}\, dt\,.$$

Replacing t by λt in this integral gives $\Gamma(s) = \lambda^s \int_0^\infty t^{s-1}\, e^{-\lambda t}\, dt$ or $\lambda^{-s} = \Gamma(s)^{-1} \int_0^\infty t^{s-1}\, e^{-\lambda t}\, dt$ for any $\lambda > 0$. Applying this to $\lambda = 2\pi n$, multiplying by a_n, and summing over n, we obtain

$$(2\pi)^{-s}\,\Gamma(s)\,L(f,s) \;=\; \sum_{n=1}^{\infty} a_n \int_0^\infty t^{s-1}\, e^{-2\pi n t}\, dt \;=\; \int_0^\infty t^{s-1}\, f(it)\, dt \qquad \left(\Re(s) > \frac{k}{2} + 1\right),$$

where the interchange of integration and summation is justified by the absolute convergence. Now the fact that $f(it)$ is exponentially small for $t \to \infty$ (because f is a cusp form) and for $t \to 0$ (because $f(-1/z) = z^k f(z)$) implies that the integral converges absolutely for all $s \in \mathbb{C}$ and hence that the function

$$L^*(f,s) \;:=\; (2\pi)^{-s}\,\Gamma(s)\,L(f,s) \;=\; (2\pi)^{-s}\,\Gamma(s) \sum_{n=1}^{\infty} \frac{a_n}{n^s} \tag{45}$$

extends holomorphically from the half-plane $\Re(s) > 1 + k/2$ to the entire complex plane. The substitution $t \to 1/t$ together with the transformation equation $f(i/t) = (it)^k f(it)$ of f then gives the functional equation

$$L^*(f, k-s) \;=\; (-1)^{k/2}\, L^*(f,s) \tag{46}$$

of $L^*(f,s)$. We have proved:

Proposition 13. *Let $f = \sum_{n=1}^{\infty} a_n q^n$ be a cusp form of weight k on the full modular group. Then the L-series $L(f,s)$ extends to an entire function of s and satisfies the functional equation* (46)*, where $L^*(f,s)$ is defined by equation* (45)*.*

It is perhaps worth mentioning that, as Hecke also proved, the converse of Proposition 13 holds as well: if a_n $(n \geq 1)$ are complex numbers of polynomial growth and the function $L^*(f,s)$ defined by (45) continues analytically to the whole complex plane and satisfies the functional equation (46), then $f(z) = \sum_{n=1}^{\infty} a_n e^{2\pi i n z}$ is a cusp form of weight k on Γ_1.

4.3 Modular Forms and Algebraic Number Theory

In §3, we used the theta series $\theta(z)^2$ to determine the number of representations of any integer n as a sum of two squares. More generally, we can study the number $r(Q,n)$ of representations of n by a positive definite binary quadratic $Q(x,y) = ax^2 + bxy + cy^2$ with integer coefficients by considering the weight 1 theta series $\Theta_Q(z) = \sum_{x,y\in\mathbb{Z}} q^{Q(x,y)} = \sum_{n=0}^{\infty} r(Q,n)q^n$. This theta series depends only on the class $[Q]$ of Q up to equivalences $Q \sim Q \circ \gamma$ with $\gamma \in \Gamma_1$. We showed in §1.2 that for any $D < 0$ the number $h(D)$ of Γ_1-equivalence classes $[Q]$ of binary quadratic forms of discriminant $b^2 - 4ac = D$ is finite. If D is a fundamental discriminant (i.e., not representable as $D'r^2$ with D' congruent to 0 or 1 mod 4 and $r > 1$), then $h(D)$ equals the class number of the imaginary quadratic field $K = \mathbb{Q}(\sqrt{D})$ of discriminant D and there is a well-known bijection between the Γ_1-equivalence classes of binary quadratic forms of discriminant D and the ideal classes of K such that $r(Q,n)$ for any form Q equals w times the number $r(\mathcal{A},n)$ of integral ideals $\mathfrak{a}$ of K of norm n belonging to the corresponding ideal class $\mathcal{A}$, where w is the number of roots of unity in K (= 6 or 4 if $D = -3$ or $D = -4$ and 2 otherwise). The L-series $L(\Theta_Q, s)$ of Θ_Q is therefore w times the "partial zeta-function" $\zeta_{K,\mathcal{A}}(s) = \sum_{\mathfrak{a}\in\mathcal{A}} N(\mathfrak{a})^{-s}$. The ideal classes of K form an abelian group. If χ is a homomorphism from this group to $\mathbb{C}^*$, then the L-series $L_K(s,\chi) = \sum_{\mathfrak{a}} \chi(\mathfrak{a})/N(\mathfrak{a})^s$ (sum over all integral ideals of K) can be written as $\sum_{\mathcal{A}} \chi(\mathcal{A})\,\zeta_{K,\mathcal{A}}(s)$ (sum over all ideal classes of K) and hence is the L-series of the weight 1 modular form $f_\chi(z) = w^{-1}\sum_{\mathcal{A}} \chi(\mathcal{A})\Theta_{\mathcal{A}}(z)$. On the other hand, from the unique prime decomposition of ideals in K it follows that $L_K(s,\chi)$ has an Euler product. Hence f_χ is a Hecke eigenform. If $\chi = \chi_0$ is the trivial character, then $L_K(s,\chi) = \zeta_K(s)$, the Dedekind zeta function of K, which factors as $\zeta(s)L(s,\varepsilon_D)$, the product of the Riemann zeta-function and the Dirichlet L-series of the character $\varepsilon_D(n) = \left(\dfrac{D}{n}\right)$ (Kronecker symbol). Therefore in this case we get $\sum_{[Q]} r(Q,n) = w\sum_{d|n}\varepsilon_D(d)$ (an identity known to Gauss) and correspondingly

$$f_{\chi_0}(z) \;=\; \frac{1}{w}\sum_{[Q]}\Theta_Q(z) \;=\; \frac{h(D)}{2} + \sum_{n=1}^{\infty}\Biggl(\sum_{d|n}\left(\frac{D}{d}\right)\Biggr)q^n\,,$$

an Eisenstein series of weight 1. If the character χ has order 2, then it is a so-called genus character and one knows that $L_K(s,\chi)$ factors as $L(s,\varepsilon_{D_1})L(s,\varepsilon_{D_2})$ where D_1 and D_2 are two other discriminants with product D. In this case, too, $f_\chi(z)$ is an Eisenstein series. But in all other cases, f_χ is a cusp form and the theory of modular forms gives us non-trivial information about representations of numbers by quadratic forms.

♠ Binary Quadratic Forms of Discriminant −23

We discuss an explicit example, taken from a short and pretty article written by van der Blij in 1952. The class number of the discriminant $D = -23$ is 3, with the $\mathrm{SL}(2,\mathbb{Z})$-equivalence classes of binary quadratic forms of this discriminant being represented by the three forms

$$\begin{aligned} Q_0(x,y) &= x^2+xy+6y^2\,,\\ Q_1(x,y) &= 2x^2+xy+3y^2\,,\\ Q_2(x,y) &= 2x^2-xy+3y^2\,. \end{aligned}$$

Since Q_1 and Q_2 represent the same integers, we get only two distinct theta series

$$\begin{aligned} \Theta_{Q_0}(z) &= 1+2q+2q^4+4q^6+4q^8+\cdots\,,\\ \Theta_{Q_1}(z) &= 1+2q^2+2q^3+2q^4+2q^6+\cdots\,. \end{aligned}$$

The linear combination corresponding to the trivial character is the Eisenstein series

$$\begin{aligned} f_{\chi_0} &= \frac{1}{2}\bigl(\Theta_{Q_0}+2\Theta_{Q_1}\bigr) \;=\; \frac{3}{2} + \sum_{n=1}^{\infty}\Biggl(\sum_{d|n}\left(\frac{-23}{d}\right)\Biggr)q^n\\ &= \frac{3}{2}+q+2q^2+2q^3+3q^4+\cdots\,, \end{aligned}$$

in accordance with the general identity $w^{-1}\sum_{[Q]} r(Q,n) = \sum_{d|n}\varepsilon_D(d)$ mentioned above. If χ is one of the two non-trivial characters, with values $e^{\pm 2\pi i/3} = \frac{1}{2}(-1\pm i\sqrt{3})$ on Q_1 and Q_2, we have

$$f_\chi = \frac{1}{2}\bigl(\varTheta_{Q_0} - \varTheta_{Q_1}\bigr) = q - q^2 - q^3 + q^6 + \cdots .$$

This is a Hecke eigenform in the space $S_1(\varGamma_0(23), \varepsilon_{-23})$. Its L-series has the form

$$L(f_\chi, s) = \prod_p \frac{1}{1 - a_p\, p^{-s} + \varepsilon_{-23}(p)\, p^{-2s}}$$

where $\varepsilon_{-23}(p)$ equals the Legendre symbol $(p/23)$ by quadratic reciprocity and

$$a_p = \begin{cases} 1 & \text{if } p = 23\,, \\ 0 & \text{if } (p/23) = -1\,, \\ 2 & \text{if } (p/23) = 1 \text{ and } p \text{ is representable as } x^2 + xy + 6y^2\,, \\ -1 & \text{if } (p/23) = 1 \text{ and } p \text{ is representable as } 2x^2 + xy + 3y^2\,. \end{cases} \tag{47}$$

On the other hand, the space $S_1(\varGamma_0(23), \varepsilon_{-23})$ is one-dimensional, spanned by the function

$$\eta(z)\,\eta(23z) = q \prod_{n=1}^{\infty} \bigl(1 - q^n\bigr)\bigl(1 - q^{23n}\bigr)\,. \tag{48}$$

We therefore obtain the explicit “reciprocity law”

Proposition 14 (van der Blij). *Let p be a prime. Then the number a_p defined in* (47) *is equal to the coefficient of q^p in the product* (48).

As an application of this, we observe that the q-expansion on the right-hand side of (48) is congruent modulo 23 to $\Delta(z) = q\prod(1-q^n)^{24}$ and hence that $\tau(p)$ is congruent modulo 23 to the number a_p defined in (47) for every prime number p, a congruence for the Ramanujan function $\tau(n)$ of a somewhat different type than those already given in (27) and (28). ♡

Proposition 14 gives a concrete example showing how the coefficients of a modular form – here $\eta(z)\eta(23z)$ – can answer a question of number theory – here, the question whether a given prime number which splits in $\mathbb{Q}(\sqrt{-23})$ splits into principal or non-principal ideals. But actually the connection goes much deeper. By elementary algebraic number theory we have that the L-series $L(s) = L_K(s,\chi)$ is the quotient $\zeta_F(s)/\zeta(s)$ of the Dedekind function of F by the Riemann zeta function, where $F = \mathbb{Q}(\alpha)$ $(\alpha^3 - \alpha - 1 = 0)$ is the cubic field of discriminant -23. (The composite $K\cdot F$ is the Hilbert class field of K.) Hence the four cases in (47) also describe the splitting of p in F: 23 is ramified, quadratic non-residues of 23 split as $p = \mathfrak{p}_1\mathfrak{p}_2$ with $N(\mathfrak{p}_i) = p^i$, and quadratic residues of 23 are either split completely (as products of three prime ideals of norm p) or are inert (remain prime) in F, according whether they are represented by Q_0 or Q_1. Thus the modular form $\eta(z)\eta(23z)$ describes not only the algebraic number theory of the quadratic field K, but also the splitting of primes in the higher degree field F. This is the first non-trivial example of the connection found by Weil–Langlands and Deligne–Serre which relates modular forms of weight one to the arithmetic of number fields whose Galois groups admit non-trivial two-dimensional representations.

4.4 Modular Forms Associated to Elliptic Curves and Other Varieties

If X is a smooth projective algebraic variety defined over $\mathbb{Q}$, then for almost all primes p the equations defining X can be reduced modulo p to define a smooth variety X_p over the field $\mathbb{Z}/p\mathbb{Z} = \mathbb{F}_p$. We can then count the number of points in X_p over the finite field $\mathbb{F}_{p^r}$ for all $r \geq 1$ and, putting all this information together, define a "local zeta function" $Z(X_p, s) = \exp\bigl(\sum_{r=1}^{\infty} |X_p(\mathbb{F}_{p^r})|\, p^{-rs}/r\bigr)$ $(\Re(s) \gg 0)$ and a "global zeta function" (Hasse–Weil zeta function) $Z(X/\mathbb{Q}, s) = \prod_p Z(X_p, s)$, where the product is over all primes. (The factors $Z_p(X,s)$ for the "bad" primes p, where the equations defining X yield a singular variety over $\overline{\mathbb{F}_p}$, are defined in a more complicated but completely explicit way and are again power series in p^{-s}.) Thanks to the work of Weil, Grothendieck, Dwork, Deligne and others, a great deal is known about the local zeta functions – in particular, that they are rational functions of p^{-s} and have all of their zeros and poles on the vertical lines $\Re(s) = 0, \frac{1}{2}, \dots, n - \frac{1}{2}, n$ where n is the dimension of X – but the global zeta function remains mysterious. In particular, the general conjecture that $Z(X/\mathbb{Q}, s)$ can be meromorphically continued to all s is known only for very special classes of varieties.

In the case where $X = E$ is an elliptic curve, given, say, by a Weierstrass equation

$$y^2 \;=\; x^3 + Ax + B \qquad (A,\, B \in \mathbb{Z}, \quad \Delta \;:=\; -4A^3 - 27B^2 \neq 0)\,, \tag{49}$$

the local factors can be made completely explicit and we find that $Z(E/\mathbb{Q}, s) = \dfrac{\zeta(s)\zeta(s-1)}{L(E/\mathbb{Q}, s)}$ where the L-series $L(E/\mathbb{Q}, s)$ is given for $\Re(s) \gg 0$ by an Euler product of the form

$$L(E/\mathbb{Q}, s) \;=\; \prod_{p \nmid \Delta} \frac{1}{1 - a_p(E)\, p^{-s} + p^{1-2s}} \cdot \prod_{p \mid \Delta} \frac{1}{(\text{polynomial of degree} \leq 2 \text{ in } p^{-s})} \tag{50}$$

with $a_p(E)$ defined for $p \nmid \Delta$ as $p - \bigl|\bigl\{(x,y) \in (\mathbb{Z}/p\mathbb{Z})^2 \mid y^2 = x^3 + Ax + B\bigr\}\bigr|$. In the mid-1950's, Taniyama noticed the striking formal similarity between this Euler product expansion and the one in (44) when $k = 2$ and asked whether there might be cases of overlap between the two, i.e., cases where the L-function of an elliptic curve agrees with that of a Hecke eigenform of weight 2 having eigenvalues $a_p \in \mathbb{Z}$ (a necessary condition if they are to agree with the integers $a_p(E)$).

Numerical examples show that this at least sometimes happens. The simplest elliptic curve (if we order elliptic curves by their "conductor," an invariant in $\mathbb{N}$ which is divisible only by primes dividing the discriminant Δ in (49)) is the curve $Y^2 - Y = X^3 - X^2$ of conductor 11. (This can be put into the form (49) by setting $y = 216Y - 108$, $x = 36X - 12$, giving $A = -432$,

$B = 8208$, $\Delta = -2^8 \cdot 3^{12} \cdot 11$, but the equation in X and Y, the so-called "minimal model," has much smaller coefficients.) We can compute the numbers a_p by counting solutions of $Y^2 - Y = X^3 - X^2$ in $(\mathbb{Z}/p\mathbb{Z})^2$. (For $p > 3$ this is equivalent to the recipe given above because the equations relating (x, y) and (X, Y) are invertible in characteristic p, and for $p = 2$ or 3 the minimal model gives the correct answer.) For example, we have $a_5 = 5 - 4 = 1$ because the equation $Y^2 - Y = X^3 - X^2$ has the 4 solutions (0,0), (0,1), (1,0) and (1,1) in $(\mathbb{Z}/5\mathbb{Z})^2$. Then we have

$$\begin{aligned} L(E/\mathbb{Q}, s) &= \left(1 + \frac{2}{2^s} + \frac{2}{2^{2s}}\right)^{-1} \left(1 + \frac{1}{3^s} + \frac{3}{3^{2s}}\right)^{-1} \left(1 - \frac{1}{5^s} + \frac{5}{5^{2s}}\right)^{-1} \cdots \\ &= \frac{1}{1^s} - \frac{2}{2^s} - \frac{1}{3^s} + \frac{2}{4^s} + \frac{1}{5^s} + \cdots = L(f, s), \end{aligned}$$

where $f \in S_2(\Gamma_0(11))$ is the modular form

$$f(z) = \eta(z)^2 \eta(11z)^2 = q \prod_{n=1}^{\infty} \bigl(1 - q^n\bigr)^2 \bigl(1 - q^{11n}\bigr)^2 = x - 2q^2 - q^3 + 2q^4 + q^5 + \cdots .$$

In the 1960's, one direction of the connection suggested by Taniyama was proved by M. Eichler and G. Shimura, whose work establishes the following theorem.

Theorem (Eichler–Shimura). *Let $f(z)$ be a Hecke eigenform in $S_2(\Gamma_0(N))$ for some $N \in \mathbb{N}$ with integral Fourier coefficients. Then there exists an elliptic curve $E/\mathbb{Q}$ such that $L(E/\mathbb{Q}, s) = L(f, s)$.*

Explicitly, this means that $a_p(E) = a_p(f)$ for all primes p, where $a_n(f)$ is the coefficient of q^n in the Fourier expansion of f. The proof of the theorem is in a sense quite explicit. The quotient of the upper half-plane by $\Gamma_0(N)$, compactified appropriately by adding a finite number of points (cusps), is a complex curve (Riemann surface), traditionally denoted by $X_0(N)$, such that the space of holomorphic 1-forms on $X_0(N)$ can be identified canonically (via $f(z) \mapsto f(z)dz$) with the space of cusp forms $S_2(\Gamma_0(N))$. One can also associate to $X_0(N)$ an abelian variety, called its *Jacobian*, whose tangent space at any point can be identified canonically with $S_2(\Gamma_0(N))$. The Hecke operators T_p introduced in 4.1 act not only on $S_2(\Gamma_0(N))$, but on the Jacobian variety itself, and if the Fourier coefficients $a_p = a_p(f)$ are in $\mathbb{Z}$, then so do the differences $T_p - a_p \cdot \mathrm{Id}$. The subvariety of the Jacobian annihilated by all of these differences (i.e., the set of points x in the Jacobian whose image under T_p equals a_p times x; this makes sense because an abelian variety has the structure of a group, so that we can multiply points by integers) is then precisely the sought-for elliptic curve E. Moreover, this construction shows that we have an even more intimate relationship between the curve E and the form f than the L-series equality $L(E/\mathbb{Q}, s) = L(f, s)$, namely, that there is an actual map from the modular curve $X_0(N)$ to the elliptic curve E which is induced by f. Specifically, if we define $\phi(z) = \sum_{n=1}^{\infty} \frac{a_n(f)}{n} e^{2\pi i n z}$, so that $\phi'(z) = 2\pi i f(z) dz$, then the fact that f is modular of weight 2 implies that the difference $\phi(\gamma(z)) - \phi(z)$ has zero derivative and hence is constant for all $\gamma \in \Gamma_0(N)$, say $\phi(\gamma(z)) - \phi(z) = C(\gamma)$. It is then easy to see that the map $C : \Gamma_0(N) \to \mathbb{C}$ is a homomorphism, and in our case (f an eigenform, eigenvalues in $\mathbb{Z}$), it turns out that its image is a lattice $\Lambda \subset \mathbb{C}$, and the quotient map $\mathbb{C}/\Lambda$ is isomorphic to the elliptic curve E. The fact that $\phi(\gamma(z)) - \phi(z) \in \Lambda$ then implies that the composite map $\mathfrak{H} \xrightarrow{\phi} \mathbb{C} \xrightarrow{\mathrm{pr}} \mathbb{C}/\Lambda$ factors through the projection $\mathfrak{H} \to \Gamma_0(N)\backslash\mathfrak{H}$, i.e., ϕ induces a map (over $\mathbb{C}$) from $X_0(N)$ to E. This map is in fact defined over $\mathbb{Q}$, i.e., there are modular functions $X(z)$ and $Y(z)$ with rational Fourier coefficients which are invariant under $\Gamma_0(N)$ and which identically satisfy the equation $Y(z)^2 = X(z)^3 + AX(z) + B$ (so that the map from $X_0(N)$ to E in its Weierstrass form is simply $z \mapsto (X(z), Y(z))$) as well as the equation $X'(z)/2Y(z) = 2\pi i f(z)$. (Here we are simplifying a little.)

Gradually the idea arose that perhaps the answer to Taniyama's original question might be yes in *all* cases, not just sometimes. The results of Eichler and Shimura showed this in one direction, and strong evidence in the other direction was provided by a theorem proved by A. Weil in 1967 which said that if the L-series of an elliptic curve $E/\mathbb{Q}$ and certain "twists" of it satisfied the conjectured analytic properties (holomorphic continuation and functional equation), then E really did correspond to a modular form in the above way. The conjecture that every E over $\mathbb{Q}$ is modular became famous (and was called according to taste by various subsets of the names Taniyama, Weil and Shimura, although none of these three people had ever stated the conjecture explicitly in print). It was finally proved at the end of the 1990's by Andrew Wiles and his collaborators and followers:

Theorem (Wiles–Taylor, Breuil–Conrad–Diamond–Taylor). *Every elliptic curve over $\mathbb{Q}$ can be parametrized by modular functions.*

The proof, which is extremely difficult and builds on almost the entire apparatus built up during the previous decades in algebraic geometry, representation theory and the theory of automorphic forms, is one of the pinnacles of mathematical achievement in the 20th century.

♠ Fermat's Last Theorem

In the 1970's, Y. Hellegouarch was led to consider the elliptic curve (49) in the special case when the roots of the cubic polynomial on the right were nth powers of rational integers for some prime number $n > 2$, i.e., if this cubic factors as $(x - a^n)(x - b^n)(x - c^n)$ where a, b, c satisfy the Fermat equation $a^n + b^n + c^n = 0$. A decade later, G. Frey studied the same elliptic curve and discovered that the associated Galois representation (we do not explain this here) had properties which contradicted the properties which Galois representations of elliptic curves were expected to satisfy. Precise conjectures about the modularity of certain Galois representations were then made by Serre which would fail for the representations attached to the Hellegouarch-Frey curve, so that the correctness of these conjectures would imply the insolubility of Fermat's equation. (Very roughly, the conjectures imply that, if the Galois representation associated to the above curve E is modular at all, then the corresponding cusp form would have to be congruent modulo n to a cusp form of weight 2 and level 1 or 2, and there aren't any.) In 1990, K. Ribet proved a special case of Serre's conjectures (the general case is now also known, thanks to recent work of Khare, Wintenberger, Dieulefait and Kisin) which was sufficient to yield the same implication. The proof by Wiles and Taylor of the Taniyama-Weil conjecture (still with some minor restrictions on E which were later lifted by the other authors cited above, but in sufficient generality to make Ribet's result applicable) thus sufficed to give the proof of the following theorem, first claimed by Fermat in 1637:

Theorem (Ribet, Wiles–Taylor). *If $n > 2$, there are no positive integers with $a^n + b^n = c^n$.* $\heartsuit$

Finally, we should mention that the connection between modularity and algebraic geometry does not apply only to elliptic curves. Without going into detail, we mention only that the Hasse–Weil zeta function $Z(X/\mathbb{Q}, s)$ of an arbitrary smooth projective variety X over $\mathbb{Q}$ splits into factors corresponding to the various cohomology groups of X, and that if any of these cohomology groups (or any piece of them under some canonical decomposition, say with respect to the action of a finite group of automorphisms of X) is two-dimensional, then the corresponding piece of the zeta function is conjectured to be the L-series of a Hecke eigenform of weight $i + 1$, where the cohomology group in question is in degree i. This of course includes the case when $X = E$ and $i = 1$, since the first cohomology group of a curve of genus 1 is 2-dimensional, but it also applies to many higher-dimensional varieties. Many examples are now known, an early one, due to R. Livné, being given by the cubic hypersurface $x_1^3 + \cdots + x_{10}^3 = 0$ in the projective space $\{x \in \mathbb{P}^9 \mid x_1 + \cdots + x_{10} = 0\}$, whose zeta-function equals $\prod_{j=0}^{7} \zeta(s - i)^{m_i} \cdot L(s - 2, f)^{-1}$ where $(m_0, \dots m_7) = (1, 1, 1, -83, 43, 1, 1, 1)$ and $f = q + 2q^2 - 8q^3 + 4q^4 + 5q^5 + \cdots$ is the unique new form of weight 4 on $\Gamma_0(10)$. Other examples arise from so-called "rigid Calabi-Yau 3-folds," which have been studied intensively in recent years in connection with the phenomenon, first discovered by mathematical physicists, called "mirror symmetry." We skip all further discussion, referring to the survey paper and book cited in the references at the end of these notes.

5 Modular Forms and Differential Operators

The starting point for this section is the observation that the derivative of a modular form is not modular, but nearly is. Specifically, if f is a modular form of weight k with the Fourier expansion (3), then by differentiating (2) we see that the derivative

$$Df \;=\; f' \;:=\; \frac{1}{2\pi i}\,\frac{df}{dz} \;=\; q\,\frac{df}{dq} \;=\; \sum_{n=1}^{\infty} n\,a_n\,q^n \tag{51}$$

(where the factor $2\pi i$ has been included in order to preserve the rationality properties of the Fourier coefficients) satisfies

$$f'\left(\frac{az+b}{cz+d}\right) \;=\; (cz+d)^{k+2}\,f'(z) \;+\; \frac{k}{2\pi i}\,c\,(cz+d)^{k+1}f(z)\,. \tag{52}$$

If we had only the first term, then f' would be a modular form of weight $k+2$. The presence of the second term, far from being a problem, makes the theory much richer. To deal with it, we will:

- modify the differentiation operator so that it preserves modularity;
- make combinations of derivatives of modular forms which are again modular;
- relax the notion of modularity to include functions satisfying equations like (52);
- differentiate with respect to $t(z)$ rather than z itself, where $t(z)$ is a modular function.

These four approaches will be discussed in the four subsections 5.1–5.4, respectively.

5.1 Derivatives of Modular Forms

As already stated, the first approach is to introduce modifications of the operator D which do preserve modularity. There are two ways to do this, one holomorphic and one not. We begin with the holomorphic one. Comparing the transformation equation (52) with equations (19) and (17), we find that for any modular form $f \in M_k(\Gamma_1)$ the function

$$\vartheta_k f \;:=\; f' - \frac{k}{12}\,E_2\,f\,, \tag{53}$$

sometimes called the *Serre derivative*, belongs to $M_{k+2}(\Gamma_1)$. (We will often drop the subscript k, since it must always be the weight of the form to which the operator is applied.) A first consequence of this basic fact is the following. We introduce the ring $\widetilde{M}_*(\Gamma_1) := M_*(\Gamma_1)[E_2] = \mathbb{C}[E_2, E_4, E_6]$, called the *ring of quasimodular forms on* $SL(2,\mathbb{Z})$. (An intrinsic definition of the elements of this ring, and a definition for other groups $\Gamma \subset G$, will be given in the next subsection.) Then we have:

Proposition 15. *The ring $\widetilde{M}_*(\Gamma_1)$ is closed under differentiation. Specifically, we have*

$$E_2' = \frac{E_2^2 - E_4}{12}, \qquad E_4' = \frac{E_2E_4 - E_6}{3}, \qquad E_6' = \frac{E_2E_6 - E_4^2}{2}. \tag{54}$$

Proof. Clearly ϑE_4 and ϑE_6, being holomorphic modular forms of weight 6 and 8 on Γ_1, respectively, must be proportional to E_6 and E_4^2, and by looking at the first terms in their Fourier expansion we find that the factors are $-1/3$ and $-1/2$. Similarly, by differentiating (19) we find the analogue of (53) for E_2, namely that the function $E_2' - \frac{1}{12}E_2^2$ belongs to $M_4(\Gamma)$. It must therefore be a multiple of E_4, and by looking at the first term in the Fourier expansion one sees that the factor is $-1/12$.

Proposition 15, first discovered by Ramanujan, has many applications. We describe two of them here. Another, in transcendence theory, will be mentioned in Section 6.

♠ Modular Forms Satisfy Non-Linear Differential Equations

An immediate consequence of Proposition 15 is the following:

Proposition 16. *Any modular form or quasi-modular form on Γ_1 satisfies a non-linear third order differential equation with constant coefficients.*

Proof. Since the ring $\widetilde{M}_*(\Gamma_1)$ has transcendence degree 3 and is closed under differentiation, the four functions f, f', f'' and f''' are algebraically dependent for any $f \in \widetilde{M}_*(\Gamma_1)$.

As an example, by applying (54) several times we find that the function E_2 satisfies the non-linear differential equation $f''' - ff'' + \frac{3}{2}f'^2 = 0$. This is called the *Chazy equation* and plays a role in the theory of Painlevé equations. We can now use modular/quasimodular ideas to describe a full set of solutions of this equation. First, define a "modified slash operator" $f \mapsto f\|_2 g$ by $(f\|_2 g)(z) = (cz+d)^{-2}f\bigl(\frac{az+b}{cz+d}\bigr) + \frac{\pi}{12}\frac{c}{cz+d}$ for $g = \bigl(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\bigr)$. This is not linear in f (it is only affine), but it is nevertheless a group operation, as one checks easily, and, at least locally, it makes sense for any matrix $\bigl(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\bigr) \in \mathrm{SL}(2,\mathbb{C})$. Now one checks by a direct, though tedious, computation that $\mathsf{Ch}\bigl[f\|_2 g\bigr] = \mathsf{Ch}[f]|_8\, g$ (where defined) for any $g \in \mathrm{SL}(2,\mathbb{C})$, where $\mathsf{Ch}[f] = f''' - ff'' + \frac{3}{2}f'^2$. (Again this is surprising, because the operator Ch is not linear.) Since E_2 is a solution of $\mathsf{Ch}[f] = 0$, it follows that $E_2\|_2 g$ is a solution of the same equation (in $g^{-1}\mathfrak{H} \subset \mathbb{P}^1(\mathbb{C})$) for every $g \in \mathrm{SL}(2,\mathbb{C})$, and since $E_2\|_2\gamma = E_2$ for $\gamma \in \Gamma_1$ and $\|_2$ is a group operation, it follows that this function depends only on the class of g in $\Gamma_1\backslash\mathrm{SL}(2,\mathbb{C})$. But $\Gamma_1\backslash\mathrm{SL}(2,\mathbb{C})$ is 3-dimensional and a third-order differential equation generically has a 3-dimensional solution space (the values of f, f' and f'' at a point determine all higher derivatives recursively and hence fix the function uniquely in a neighborhood of any point where it is holomorphic), so that, at least generically, this describes all solutions of the non-linear differential equation $\mathsf{Ch}[f] = 0$ in modular terms. ♡

Our second "application" of the ring $\widetilde{M}_*(\Gamma_1)$ describes an unexpected appearance of this ring in an elementary and apparently unrelated context.

♠ Moments of Periodic Functions

This very pretty application of the modular forms E_2, E_4, E_6 is due to P. Gallagher. Denote by $\mathfrak{P}$ the space of periodic real-valued functions on $\mathbb{R}$, i.e., functions $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(x + 2\pi) = f(x)$. For each ∞-tuple $\mathbf{n} = (n_0, n_1, \dots)$ of non-negative integers (all but finitely many equal to 0) we define a "coordinate" $I_{\mathbf{n}}$ on the infinite-dimensional space $\mathfrak{P}$ by $I_{\mathbf{n}}[f] = \int_0^{2\pi} f(x)^{n_0} f'(x)^{n_1} \cdots dx$ (higher moments). Apart from the relations among these coming from integration by parts, like $\int f''(x)dx = 0$ or $\int f'(x)^2 dx = -\int f(x)f''(x)dx$, we also have various inequalities. The general problem, certainly too hard to be solved completely, would be to describe all equalities and inequalities among the $I_{\mathbf{n}}[f]$. As a special case we can ask for the complete list of inequalities satisfied by the four moments $\bigl(A[f], B[f], C[f], D[f]\bigr) := \bigl(\int_0^{2\pi} f,\ \int_0^{2\pi} f^2,\ \int_0^{2\pi} f^3,\ \int_0^{2\pi} f'^2\bigr)$ as f ranges over $\mathfrak{P}$. Surprisingly enough, the answer involves quasimodular forms on $\mathrm{SL}(2,\mathbb{Z})$. First, by making a linear shift $f \mapsto \lambda f + \mu$ with λ, $\mu \in \mathbb{R}$ we can suppose that $A[f] = 0$ and $D[f] = 1$. The problem is then to describe the subset $\mathfrak{X} \subset \mathbb{R}^2$ of pairs $\bigl(B[f],\, C[f]\bigr) = \bigl(\int_0^{2\pi} f(x)^2 dx,\ \int_0^{2\pi} f(x)^3 dx\bigr)$ where f ranges over functions in $\mathfrak{P}$ satisfying $\int_0^{2\pi} f(x)dx = 0$ and $\int_0^{2\pi} f'(x)^2\, dx = 1$.

Theorem (Gallagher). *We have* $\mathfrak{X} = \bigl\{\, (B, C) \in \mathbb{R}^2 \mid 0 < B \le 1,\ C^2 \le \Phi(B) \,\bigr\}$ *where the function* $\Phi : (0, 1] \to \mathbb{R}_{\ge 0}$ *is given parametrically by*

$$\Phi\biggl(\frac{\mathbb{G}_2'(it)}{\mathbb{G}_4'(it)}\biggr) = \frac{(\mathbb{G}_4'(it) - \mathbb{G}_2''(it))^2}{2\,\mathbb{G}_4'(it)^3} \qquad \bigl(0 < t \le \infty\bigr)\,.$$

The idea of the proof is as follows. First, from $A = 0$ and $D = 1$ we deduce $0 < B[f] \le 1$ by an inequality of Wirtinger (just look at the Fourier expansion of f). Now let $f \in \mathfrak{P}$ be a function – but one must prove that it exists! – which maximizes $C = C[f]$ for given values of A, B and D. By a standard calculus-of-variations-type argument (replace f by $f + \varepsilon g$ where $g \in \mathfrak{P}$ is orthogonal to 1, f and f'', so that A, B and D do not change to first order, and then use that C also cannot change to first order since otherwise its value could not be extremal, so that g must also be orthogonal to f^2), we show that the four functions 1, f, f^2 and f'' are linearly dependent. From this it follows by integrating once that the five functions 1, f, f^2, f^3 and f'^2 are also linearly dependent. After a rescaling $f \mapsto \lambda f + \mu$, we can write this dependency as $f'(x)^2 = 4f(x)^3 - g_2 f(x) - g_3$ for some constants g_2 and g_3. But this is the famous differential equation of the Weierstrass $\wp$-function, so $f(2\pi x)$ is the restriction to $\mathbb{R}$ of the function $\wp(x, \mathbb{Z}\tau + \mathbb{Z})$ for some $\tau \in \mathfrak{H}$, necessarily of the form $\tau = it$ with $t > 0$ because everything is real. Now the coefficients g_2 and g_3, by the classical Weierstrass theory, are simple multiples of $\mathbb{G}_4(it)$ and $\mathbb{G}_6(it)$, and for this function f the value of $A(f) = \int_0^{2\pi} f(x)dx$ is known to be a multiple of $\mathbb{G}_2(it)$. Working out all the details, and then rescaling f again to get $A = 0$ and $D = 1$, one finds the result stated in the theorem. ♡

We now turn to the second modification of the differentiation operator which preserves modularity, this time, however, at the expense of sacrificing holomorphy. For $f \in M_k(\Gamma)$ (we now no longer require that Γ be the full modular group Γ_1) we define

$$\partial_k f(z) \;=\; f'(z) - \frac{k}{4\pi y} f(z)\,, \tag{55}$$

where y denotes the imaginary part of z. Clearly this is no longer holomorphic, but from the calculation

$$\frac{1}{\Im(\gamma z)} \;=\; \frac{|cz+d|^2}{y} \;=\; \frac{(cz+d)^2}{y} - 2ic(cz+d) \qquad \left(\gamma \;=\; \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{R})\right)$$

and (52) one easily sees that it transforms like a modular form of weight $k+2$, i.e., that $(\partial_k f)|_{k+2}\gamma = \partial_k f$ for all $\gamma \in \Gamma$. Moreover, this remains true even if f is modular but not holomorphic, if we interpret f' as $\dfrac{1}{2\pi i}\dfrac{\partial f}{\partial z}$. This means that we can apply $\partial = \partial_k$ repeatedly to get non-holomorphic modular forms $\partial^n f$ of weight $k + 2n$ for all $n \geq 0$. (Here, as with ϑ_k, we can drop the subscript k because ∂_k will only be applied to forms of weight k; this is convenient because we can then write $\partial^n f$ instead of the more correct $\partial_{k+2n-2} \cdots \partial_{k+2}\partial_k f$.) For example, for $f \in M_k(\Gamma)$ we find

$$\begin{aligned}
\partial^2 f \;&=\; \left(\frac{1}{2\pi i}\frac{\partial}{\partial z} - \frac{k+2}{4\pi y}\right)\left(f' - \frac{k}{4\pi y} f\right) \\
&=\; f'' - \frac{k}{4\pi y} f' - \frac{k}{16\pi^2 y^2} f - \frac{k+2}{4\pi y} f' + \frac{k(k+2)}{16\pi^2 y^2} f \\
&=\; f'' - \frac{k+1}{2\pi y} f' + \frac{k(k+1)}{16\pi^2 y^2} f
\end{aligned}$$

and more generally, as one sees by an easy induction,

$$\partial^n f \;=\; \sum_{r=0}^{n} (-1)^{n-r} \binom{n}{r} \frac{(k+r)_{n-r}}{(4\pi y)^{n-r}} D^r f\,, \tag{56}$$

where $(a)_m = a(a+1)\cdots(a+m-1)$ is the Pochhammer symbol. The inversion of (56) is

$$D^n f \;=\; \sum_{r=0}^{n} \binom{n}{r} \frac{(k+r)_{n-r}}{(4\pi y)^{n-r}}\, \partial^r f\,, \tag{57}$$

and describes the decomposition of the holomorphic but non-modular form $f^{(n)} = D^n f$ into non-holomorphic but modular pieces: the function $y^{r-n}\partial^r f$ is multiplied by $(cz+d)^{k+n+r}(c\bar{z}+d)^{n-r}$ when z is replaced by $\frac{az+b}{cz+d}$ with $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$.

Formula (56) has a consequence which will be important in §6. The usual way to write down modular forms is via their Fourier expansions, i.e., as power series in the quantity $q = e^{2\pi i z}$ which is a local coordinate at infinity for the modular curve $\Gamma\backslash\mathfrak{H}$. But since modular forms are holomorphic functions in the upper half-plane, they also have Taylor series expansions in the neighborhood of any point $z = x + iy \in \mathfrak{H}$. The "straight" Taylor series expansion, giving $f(z+w)$ as a power series in w, converges only in the disk $|w| < y$ centered at z and tangent to the real line, which is unnatural since the domain of holomorphy of f is the whole upper half-plane, not just this disk. Instead, we should remember that we can map $\mathfrak{H}$ isomorphically to the unit disk, with z mapping to 0, by sending $z' \in \mathfrak{H}$ to $w = \frac{z'-z}{z'-\bar{z}}$. The inverse of this map is given by $z' = \frac{z-\bar{z}w}{1-w}$, and then if f is a modular form of weight k we should also include the automorphy factor $(1-w)^{-k}$ corresponding to this fractional linear transformation (even though it belongs to $\mathrm{PSL}(2,\mathbb{C})$ and not Γ). The most natural way to study f near z is therefore to expand $(1-w)^{-k} f\bigl(\frac{z-\bar{z}w}{1-w}\bigr)$ in powers of w. The following proposition describes the coefficients of this expansion in terms of the operator (55).

Proposition 17. *Let f be a modular form of weight k and $z = x+iy$ a point of $\mathfrak{H}$. Then*

$$\bigl(1-w\bigr)^{-k}\, f\biggl(\frac{z-\bar{z}w}{1-w}\biggr) \;=\; \sum_{n=0}^{\infty} \partial^n f(z)\, \frac{(4\pi y w)^n}{n!} \qquad (|w| < 1)\,. \tag{58}$$

Proof. From the usual Taylor expansion, we find

$$\begin{aligned} \bigl(1-w\bigr)^{-k} f\biggl(\frac{z-\bar{z}w}{1-w}\biggr) \;&=\; \bigl(1-w\bigr)^{-k} f\biggl(z + \frac{2iyw}{1-w}\biggr) \\ &=\; \bigl(1-w\bigr)^{-k} \sum_{r=0}^{\infty} \frac{D^r f(z)}{r!} \biggl(\frac{-4\pi y w}{1-w}\biggr)^{r}, \end{aligned}$$

and now expanding $(1-w)^{-k-r}$ by the binomial theorem and using (56) we obtain (58).

Proposition 17 is useful because, as we will see in §6, the expansion (58), after some renormalizing, often has algebraic coefficients that contain interesting arithmetic information.

5.2 Rankin–Cohen Brackets and Cohen–Kuznetsov Series

Let us return to equation (52) describing the near-modularity of the derivative of a modular form $f \in M_k(\Gamma)$. If $g \in M_\ell(\Gamma)$ is a second modular form on the same group, of weight ℓ, then this formula shows that the non-modularity of $f'(z)g(z)$ is given by an additive correction term $(2\pi i)^{-1}kc\,(cz+d)^{k+\ell+1}f(z)\,g(z)$. This correction term, multiplied by ℓ, is symmetric in f and g, so the difference $[f,g] = kfg' - \ell f'g$ is a modular form of weight $k+\ell+2$ on Γ. One checks easily that the bracket $[\,\cdot\,,\,\cdot\,]$ defined in this way is anti-symmetric and satisfies the Jacobi identity, making $M_*(\Gamma)$ into a graded Lie algebra (with grading given by the weight $+\,2$). Furthermore, the bracket $g \mapsto [f,g]$ with a fixed modular form f is a derivation with respect to the usual multiplication, so that $M_*(\Gamma)$ even acquires the structure of a so-called Poisson algebra.

We can continue this construction to find combinations of higher derivatives of f and g which are modular, setting $[f,g]_0 = fg$, $[f,g]_1 = [f,g] = kfg' - \ell f'g$,

$$[f,g]_2 \;=\; \frac{k(k+1)}{2}fg'' \;-\; (k+1)(\ell+1)f'g' \;+\; \frac{\ell(\ell+1)}{2}\,f''g\,,$$

and in general

$$[f,g]_n \;=\; \sum_{\substack{r,\,s\ge 0\\ r+s=n}} (-1)^r \binom{k+n-1}{s}\binom{\ell+n-1}{r} D^rf\,D^sg \qquad (n\ge 0)\,, \quad (59)$$

the nth *Rankin–Cohen bracket* of f and g.

Proposition 18. *For $f \in M_k(\Gamma)$ and $g \in M_\ell(\Gamma)$ and for every $n \ge 0$, the function $[f,g]_n$ defined by* (59) *belongs to $M_{k+\ell+2n}(\Gamma)$.*

There are several ways to prove this. We will do it using *Cohen–Kuznetsov series.* If $f \in M_k(\Gamma)$, then the Cohen–Kuznetsov series of f is defined by

$$\widetilde{f}_D(z,X) \;=\; \sum_{n=0}^{\infty} \frac{D^nf(z)}{n!\,(k)_n}\,X^n \quad \in\; \mathrm{Hol}_0(\mathfrak{H})[[X]]\,, \qquad (60)$$

where $(k)_n = (k+n-1)!/(k-1)! = k(k+1)\cdots(k+n-1)$ is the Pochhammer symbol already used above and $\mathrm{Hol}_0(\mathfrak{H})$ denotes the space of holomorphic functions in the upper half-plane of subexponential growth (see §1.1). This series converges for all $X \in \mathbb{C}$ (although for our purposes only its properties as a formal power series in X will be needed). Its key property is given by:

Proposition 19. *If $f \in M_k(\Gamma)$, then the Cohen–Kuznetsov series defined by* (60) *satisfies the modular transformation equation*

$$\widetilde{f}_D\biggl(\frac{az+b}{cz+d},\,\frac{X}{(cz+d)^2}\biggr) \;=\; (cz+d)^k\,\exp\biggl(\frac{c}{cz+d}\,\frac{X}{2\pi i}\biggr)\,\widetilde{f}_D(z,X)\,. \qquad (61)$$

for all $z \in \mathfrak{H}$, $X \in \mathbb{C}$, and $\gamma = \left(\begin{smallmatrix} a & b\\ c & d\end{smallmatrix}\right) \in \Gamma$.

Proof. This can be proved in several different ways. One way is direct: one shows by induction on n that the derivative $D^n f(z)$ transforms under Γ by

$$D^n f\left(\frac{az+b}{cz+d}\right) = \sum_{r=0}^{n} \binom{n}{r} \frac{(k+r)_{n-r}}{(2\pi i)^{n-r}} \, c^{n-r} (cz+d)^{k+n+r} \, D^r f(z)$$

for all $n \ge 0$ (equation (52) is the case $n=1$ of this), from which the claim follows easily. Another, more elegant, method is to use formula (56) or (57) to establish the relationship

$$\widetilde{f}_D(z,X) = e^{X/4\pi y} \, \widetilde{f}_\partial(z,X) \qquad (z = x+iy \in \mathfrak{H}, \ X \in \mathbb{C}) \tag{62}$$

between $\widetilde{f}_D(z,X)$ and the modified Cohen–Kuznetsov series

$$\widetilde{f}_\partial(z,X) = \sum_{n=0}^{\infty} \frac{\partial^n f(z)}{n! \, (k)_n} X^n \quad \in \operatorname{Hol}_0(\mathfrak{H})[[X]]\,. \tag{63}$$

The fact that each function $\partial^n f(z)$ transforms like a modular form of weight $k+2n$ on Γ implies that $\widetilde{f}_\partial(z,X)$ is multiplied by $(cz+d)^k$ when z and X are replaced by $\frac{az+b}{cz+d}$ and $\frac{X}{(cz+d)^2}$, and using (62) one easily deduces from this the transformation formula (61). Yet a third way is to observe that $\widetilde{f}_D(z,X)$ is the unique solution of the differential equation $\bigl(X\frac{\partial^2}{\partial X^2} + k\frac{\partial}{\partial X} - D\bigr)\widetilde{f}_D = 0$ with the initial condition $\widetilde{f}_D(z,0) = f(z)$ and that $(cz+d)^{-k} e^{-cX/2\pi i(cz+d)} \widetilde{f}_D\bigl(\frac{az+b}{cz+d}, \frac{X}{(cz+d)^2}\bigr)$ satisfies the same differential equation with the same initial condition.

Now to deduce Proposition 18 we simply look at the product of $\widetilde{f}_D(z,-X)$ with $\widetilde{g}_D(z,X)$. Proposition 19 implies that this product is multiplied by $(cz+d)^{k+\ell}$ when z and X are replaced by $\frac{az+b}{cz+d}$ and $\frac{X}{(cz+d)^2}$ (the factors involving an exponential in X cancel), and this means that the coefficient of X^n in the product, which is equal to $\frac{[f,g]_n}{(k)_n(\ell)_n}$, is modular of weight $k+\ell+2n$ for every $n \ge 0$.

Rankin–Cohen brackets have many applications in the theory of modular forms. We will describe two – one very straightforward and one more subtle – at the end of this subsection, and another one in §5.4. First, however, we make a further comment about the Cohen–Kuznetsov series attached to a modular form. We have already introduced two such series: the series $\widetilde{f}_D(z,X)$ defined by (60), with coefficients proportional to $D^n f(z)$, and the series $\widetilde{f}_\partial(z,X)$ defined by (63), with coefficients proportional to $\partial^n f(z)$. But, at least when Γ is the full modular group Γ_1, we had defined a third differentiation operator besides D and ∂, namely the operator ϑ defined in (53), and it is natural to ask whether there is a corresponding Cohen–Kuznetsov series $\widetilde{f}_\vartheta$ here also. The answer is yes, but this series is not simply given by $\sum_{n\ge 0}\vartheta^n f(z)\,X^n/n!\,(k)_n$. Instead, for $f\in M_k(\Gamma_1)$, we define a sequence of modified derivatives $\vartheta^{[n]}f\in M_{k+2n}(\Gamma_1)$ for $n\ge 0$ by

$$\vartheta^{[0]}f \;=\; f,\;\; \vartheta^{[1]}f \;=\; \vartheta f,\;\; \vartheta^{[r+1]}f \;=\; \vartheta\bigl(\vartheta^{[r]}f\bigr)-r(k+r-1)\frac{E_4}{144}\vartheta^{[r-1]}f \text{ for } r\ge 1 \tag{64}$$

(the last formula also holds for $r=0$ with $f^{[-1]}$ defined as 0 or in any other way), and set

$$\widetilde{f}_\vartheta(z,X) \;=\; \sum_{n=0}^{\infty}\frac{\vartheta^{[n]}f(z)}{n!\,(k)_n}\,X^n\,.$$

Using the first equation in (54) we find by induction on n that

$$D^n f \;=\; \sum_{r=0}^{n}\binom{n}{r}(k+r)_{n-r}\left(\frac{E_2}{12}\right)^{n-r}\vartheta^{[r]}f \qquad (n\;=\;0,1,\dots)\,, \tag{65}$$

(together with similar formulas for $\partial^n f$ in terms of $\vartheta^{[n]}f$ and for $\vartheta^{[n]}f$ in terms of $D^n f$ or $\partial^n f$), the explicit version of the expansion of $D^n f$ as a polynomial in E_2 with modular coefficients whose existence is guaranteed by Proposition 15. This formula together with (62) gives us the relations

$$\widetilde{f}_\vartheta(z,X) \;=\; e^{-XE_2(z)/12}\,\widetilde{f}_D(z,X) \;=\; e^{-XE_2^*(z)/12}\,\widetilde{f}_\partial(z,X) \tag{66}$$

between the new series and the old ones, where $E_2^*(z)$ is the non-holomorphic Eisenstein series defined in (21), which transforms like a modular form of weight 2 on Γ_1. More generally, if we are on any discrete subgroup Γ of $\mathrm{SL}(2,\mathbb{R})$, we choose a holomorphic or meromorphic function ϕ in $\mathfrak{H}$ such that the function $\phi^*(z)=\phi(z)-\frac{1}{4\pi y}$ transforms like a modular form of weight 2 on Γ, or equivalently such that $\phi\bigl(\frac{az+b}{cz+d}\bigr) \;=\; (cz+d)^2\phi(z)+\frac{1}{2\pi i}c(cz+d)$ for all $\bigl(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\bigr)\in\Gamma$. (Such a ϕ always exists, and if Γ is commensurable with Γ_1 is simply the sum of $\frac{1}{12}E_2(z)$ and a holomorphic or meromorphic modular form of weight 2 on $\Gamma\cap\Gamma_1$.) Then, just as in the special case $\phi=E_2/12$, the operator ϑ_ϕ defined by $\vartheta_\phi f:=Df-k\phi f$ for $f\in M_k(\Gamma)$ sends $M_k(\Gamma)$ to $M_{k+2}(\Gamma)$, the function $\omega:=\phi'-\phi^2$ belongs to $M_4(\Gamma)$ (generalizing the first equation in (54)), and if we generalize the above definition by introducing operators $\vartheta_\phi^{[n]}:\;M_k(\Gamma)\to M_{k+2n}(\Gamma)$ $(n\;=\;0,1,\dots)$ by

$$\vartheta_\phi^{[0]}f \;=\; f\,, \qquad \vartheta_\phi^{[r+1]}f \;=\; \vartheta_\phi^{[r]}f\,+\,r(k+r-1)\,\omega\,\vartheta_\phi^{[r-1]}f \quad \text{for } r\ge 0\,, \tag{67}$$

then (65) holds with $\frac{1}{12}E_2$ replaced by ϕ, and (66) is replaced by

$$\widetilde{f}_{\vartheta_\phi}(z,X) \;:=\; \sum_{n=0}^{\infty}\frac{\vartheta_\phi^{[n]}f(z)\,X^n}{n!\,(k)_n} \;=\; e^{-\phi(z)X}\,\widetilde{f}_D(z,X) \;=\; e^{-\phi^*(z)X}\,\widetilde{f}_\partial(z,X)\,. \tag{68}$$

These formulas will be used again in §§6.3–6.4.

We now give the two promised applications of Rankin–Cohen brackets.

♠ Further Identities for Sums of Powers of Divisors

In §2.2 we gave identities among the divisor power sums $\sigma_\nu(n)$ as one of our first applications of modular forms and of identities like $E_4^2 = E_8$. By including the quasimodular form E_2 we get many more identities of the same type, e.g., the relationship $E_2^2 = E_4 + 12E_2'$ gives the identity $\sum_{m=1}^{n-1}\sigma_1(m)\sigma_1(n-m) = \frac{1}{12}\bigl(5\sigma_3(n) - (6n-1)\sigma_1(n)\bigr)$ and similarly for the other two formulas in (54). Using the Rankin–Cohen brackets we get yet more. For instance, the first Rankin–Cohen bracket of $E_4(z)$ and $E_6(z)$ is a cusp form of weight 12 on Γ_1, so it must be a multiple of $\Delta(z)$, and this leads to the formula $\tau(n) = n(\frac{7}{12}\sigma_5(n) + \frac{5}{12}\sigma_3(n)) - 70\sum_{m=1}^{n-1}(5m-2n)\sigma_3(m)\sigma_5(n-m)$ for the coefficient $\tau(n)$ of q^n in Δ. (This can also be expressed completely in terms of elementary functions, without mentioning $\Delta(z)$ or $\tau(n)$, by writing Δ as a linear combination of any two of the three functions E_{12}, E_4E_8 and E_6^2.) As in the case of the identities mentioned in §2.2, all of these identities also have combinatorial proofs, but these involve much more work and more thought than the (quasi)modular ones. ♡

♠ Exotic Multiplications of Modular Forms

A construction which is familiar both in symplectic geometrys and in quantum theory (Moyal brackets) is that of the deformation of the multiplication in an algebra. If A is an algebra over some field k, say with a commutative and associative multiplication $\mu : A \otimes A \to A$, then we can look at deformations μ_ε of μ given by formal power series $\mu_\varepsilon(x,y) = \mu_0(x,y) + \mu_1(x,y)\varepsilon + \mu_2(x,y)\varepsilon^2 + \cdots$ with $\mu_0 = \mu$ which are still associative but are no longer necessarily commutative. (To be more precise, μ_ε is a multiplication on A if there is a topology and the series above is convergent and otherwise, after being extended $k[[\varepsilon]]$-linearly, a multiplication on $A[[\varepsilon]]$.) The linear term $\mu_1(x,y)$ in the expansion of μ_ε is then anti-symmetric and satisfies the Jacobi identity, making A (or $A[[\varepsilon]]$) into a Lie algebra, and the μ_1-product with a fixed element of A is a derivation with respect to the original multiplication μ, giving the structure of a Poisson algebra. All of this is very reminiscent of the zeroth and first Rankin–Cohen brackets, so one can ask whether these two brackets arise as the beginning of the expansion of some deformation of the ordinary multiplication of modular forms. Surprisingly, this is not only true, but there is in fact a *two-parameter* family of such deformed multiplications:

Theorem. *Let u and v be two formal variables and $\mathit{\Gamma}$ any subgroup of $SL(2,\mathbb{R})$. Then the multiplication $\mu_{u,v}$ on $\prod\limits_{k=0}^{\infty} M_k(\mathit{\Gamma})$ defined by*

$$\mu_{u,v}(f,g) \;=\; \sum_{n=0}^{\infty} t_n(k,\ell;u,v)\,[f,g]_n \qquad (f\in M_k(\mathit{\Gamma}),\ g\in M_\ell(\mathit{\Gamma}))\,, \tag{69}$$

where the coefficients $t_n(k,\ell;u,v)\in\mathbb{Q}(k,l)[u,v]$ are given by

$$t_n(k,\ell;u,v) \;=\; v^n \sum_{0\le j\le n/2} \binom{n}{2j} \frac{\binom{-\frac12}{j}\binom{\frac{u}{v}-1}{j}\binom{-\frac{u}{v}}{j}}{\binom{-k-\frac12}{j}\binom{-\ell-\frac12}{j}\binom{n+k+l-\frac32}{j}}\,, \tag{70}$$

is associative.

Of course the multiplication given by $(u,v)=(0,0)$ is just the usual multiplication of modular forms, and the multiplications associated to (u,v) and $(\lambda u,\lambda v)$ are isomorphic by rescaling $f\mapsto\lambda^k f$ for $f\in M_k$, so the set of new multiplications obtained this way is parametrized by a projective line. Two of the multiplications obtained are noteworthy: the one corresponding to $(u:v)=(1:0)$ because it is the only commutative one, and the one corresponding to $(u,v)=(0,1)$ because it is so simple, being given just by $f*g=\sum_n[f,g]_n$.

These deformed multiplications were found, at approximately the same time, in two independent investigations and as consequences of two quite different theories. On the one hand, Y. Manin, P. Cohen and I studied $\mathit{\Gamma}$-invariant and twisted $\mathit{\Gamma}$-invariant pseudodifferential operators in the upper half-plane. These are formal power series $\mathit{\Psi}(z)=\sum_{n\ge h} f_n(z)D^{-n}$, with D as in (52), transforming under $\mathit{\Gamma}$ by $\mathit{\Psi}\bigl(\frac{az+b}{cz+d}\bigr)=\mathit{\Psi}(z)$ or by $\mathit{\Psi}\bigl(\frac{az+b}{cz+d}\bigr)=(cz+d)^\kappa\mathit{\Psi}(z)(cz+d)^{-\kappa}$, respectively, where κ is some complex parameter. (Notice that the second formula is different from the first because multiplication by a non-constant function of z does not commute with the differentiation operator D, and is well-defined even for non-integral κ because the ambiguity of argument involved in choosing a branch of $(cz+d)^\kappa$ cancels out when one divides by $(cz+d)^\kappa$ on the other side.) Using that D transforms under the action of $\bigl(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\bigr)$ to $(cz+d)^2D$, one finds that the leading coefficient f_h of F is then a modular form on $\mathit{\Gamma}$ of weight $2h$ and that the higher coefficients f_{h+j} $(j\ge0)$ are specific linear combinations (depending on h, j and κ) of D^jg_0, $D^{j-1}g_1,\ldots,$ g_j for some modular forms $g_i\in M_{2h+2i}(\mathrm{G})$, so that (assuming that $\mathit{\Gamma}$ contains -1 and hence has only modular forms of even weight) we can canonically identify the space of invariant or twisted invariant pseudodifferential operators with $\prod_{k=0}^\infty M_k(\mathit{\Gamma})$. On the other hand, pseudo-differential operators can be multiplied just by multiplying out the formal series defining them using Leibnitz's rule, this multiplication clearly being associative, and this then leads to the family of non-trivial multiplications of modular forms given in the theorem, with $u/v=\kappa-1/2$. The other paper in which the same multiplications arise is one by A. and J. Untenberger which is based on a certain realization of modular forms as Hilbert-Schmidt operators on $L^2(\mathbb{R})$. In both constructions, the coefficients $t_n(k,\ell;u,v)$ come out in a different and less symmetric form than (70); the equivalence with the form given above is a very complicated combinatorial identity. ♡

5.3 Quasimodular Forms

We now turn to the definition of the ring $\widetilde{M}_*(\Gamma)$ of quasimodular forms on Γ. In §5.1 we defined this ring for the case $\Gamma = \Gamma_1$ as $\mathbb{C}[E_2, E_4, E_6]$. We now give an intrinsic definition which applies also to other discrete groups Γ.

We start by defining the ring of *almost holomorphic modular forms* on Γ. By definition, such a form is a function in $\mathfrak{H}$ which transforms like a modular form but, instead of being holomorphic, is a polynomial in $1/y$ (with $y = \Im(z)$ as usual) with holomorphic coefficients, the motivating examples being the non-holomorphic Eisenstein series $E_2^*(z)$ and the non-holomorphic derivative $\partial f(z)$ of a holomorphic modular form as defined in equations (21) and (55), respectively. More precisely, an almost holomorphic modular form of *weight* k and *depth* $\le p$ on Γ is a function of the form $F(z) = \sum_{r=0}^{p} f_r(z)(-4\pi y)^{-r}$ with each $f_r \in \mathrm{Hol}_0(\mathfrak{H})$ (holomorphic functions of moderate growth) satisfying $F|_k\gamma = F$ for all $\gamma \in \Gamma$. We denote by $\widehat{M}_k^{(\le p)} = \widehat{M}_k^{(\le p)}(\Gamma)$ the space of such forms and by $\widehat{M}_* = \bigoplus_k \widehat{M}_k$, $\widehat{M}_k = \cup_p \widehat{M}_k^{(\le p)}$ the graded and filtered ring of all almost holomorphic modular forms, usually omitting Γ from the notations. For the two basic examples $E_2^* \in \widehat{M}_2^{(\le 1)}(\Gamma_1)$ and $\partial_k f \in \widehat{M}_{k+2}^{(\le 1)}(\Gamma)$ (where $f \in M_k(\Gamma)$) we have $f_0 = E_2$, $f_1 = 12$ and $f_0 = Df$, $f_1 = kf$, respectively.

We now define the space $\widetilde{M}_k^{(\le p)} = \widetilde{M}_k^{(\le p)}(\Gamma)$ of quasimodular forms of weight k and depth $\le p$ on Γ as the space of "constant terms" $f_0(z)$ of $F(z)$ as F runs over $\widehat{M}_k^{(\le p)}$. It is not hard to see that the almost holomorphic modular form F is uniquely determined by its constant term f_0, so the ring $\widetilde{M}_* = \bigoplus_k \widetilde{M}_k$ $(\widetilde{M}_k = \cup_p \widetilde{M}_k^{(\le p)})$ of quasimodular forms on Γ is canonically isomorphic to the ring $\widehat{M}_*$ of almost holomorphic modular forms on Γ. One can also define quasimodular forms directly, as was pointed out to me by W. Nahm: a quasimodular form of weight k and depth $\le p$ on Γ is a function $f \in \mathrm{Hol}_0(\mathfrak{H})$ such that, for fixed $z \in \mathfrak{H}$ and variable $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$, the function $\bigl(f|_k\gamma\bigr)(z)$ is a polynomial of degree $\le p$ in $\frac{c}{cz+d}$. Indeed, if $f(z) = f_0(z) \in \widetilde{M}_k$ corresponds to $F(z) = \sum_r f_r(z)\,(-4\pi y)^{-r} \in \widehat{M}_k$, then the modularity of F implies the identity $\bigl(f|_k\gamma\bigr)(z) = \sum_r f_r(z)\bigl(\frac{c}{cz+d}\bigr)^r$ with the same coefficients $f_r(z)$, and conversely.

The basic facts about quasimodular forms are summarized in the following proposition, in which Γ is a non-cocompact discrete subgroup of $\mathrm{SL}(2,\mathbb{R})$ and $\phi \in \widetilde{M}_2(\Gamma)$ is a quasimodular form of weight 2 on Γ which is not modular, e.g., $\phi = E_2$ if Γ is a subgroup of Γ_1. For $\Gamma = \Gamma_1$ part (i) of the proposition reduces to Proposition 15 above, while part (ii) shows that the general definition of quasimodular forms given here agrees with the ad hoc one given in §5.1 in this case.

Proposition 20. (i) *The space of quasimodular forms on $\mathit{\Gamma}$ is closed under differentiation. More precisely, we have $D\big(\widetilde{M}_k^{(\le p)}\big) \subseteq \widetilde{M}_{k+2}^{(\le p+1)}$ for all $k,\,p \ge 0$.*
(ii) *Every quasimodular form on $\mathit{\Gamma}$ is a polynomial in ϕ with modular coefficients. More precisely, we have $\widetilde{M}_k^{(\le p)}(\mathit{\Gamma}) = \bigoplus_{r=0}^{p} M_{k-2r}(\mathit{\Gamma})\cdot\phi^r$ for all $k,\,p \ge 0$.*
(iii) *Every quasimodular form on $\mathit{\Gamma}$ can be written uniquely as a linear combination of derivatives of modular forms and of ϕ. More precisely, for all $k,\,p \ge 0$ we have*

$$\widetilde{M}_k^{(\le p)}(\mathit{\Gamma}) \;=\; \begin{cases} \bigoplus_{r=0}^{p} D^r\big(M_{k-2r}(\mathit{\Gamma})\big) & \text{if } p < k/2\,, \\ \bigoplus_{r=0}^{k/2-1} D^r\big(M_{k-2r}(\mathit{\Gamma})\big) \,\oplus\, \mathbb{C}\cdot D^{k/2-1}\phi & \text{if } p \ge k/2\,. \end{cases}$$

Proof. Let $F = \sum f_r(-4\pi y)^{-r} \in \widehat{M}_k$ correspond to $f = f_0 \in \widetilde{M}_k$. The almost holomorphic form $\partial_k F \in \widehat{M}_{k+2}$ then has the expansion $\partial_k F = \sum\big[D(f_r) + (k-r+1)f_{r-1}\big](-4\pi y)^{-r}$, with constant term Df. This proves the first statement. (One can also prove it in terms of the direct definition of quasimodular forms by differentiating the formula expressing the transformation behavior of f under $\mathit{\Gamma}$.) Next, one checks easily that if F belongs to $\widehat{M}_k^{(\le p)}$, then the last coefficient $f_p(z)$ in its expansion is a modular form of weight $k-2p$. It follows that $p \le k/2$ (if $f_p \ne 0$, i.e., if F has depth exactly p) and also, since the almost holomorphic modular form ϕ^* corresponding to ϕ is the sum of ϕ and a non-zero multiple of $1/y$, that F is a linear combination of $f_p\phi^{*p}$ and an almost holomorphic modular form of depth strictly smaller than p, from which statement (ii) for almost holomorphic modular forms (and therefore also for quasimodular forms) follows by induction on p. Statement (iii) is proved exactly the same way, by subtracting from F a multiple of $\partial^p f_p$ if $p < k/2$ and a multiple of $\partial^{k/2-1}\phi$ if $p = k/2$ to prove by induction on p the corresponding statement with "quasimodular" and D replaced by "almost holomorphic modular" and ∂, and then again using the isomorphism between $\widehat{M}_*$ and $\widetilde{M}_*$.

There is one more important element of the structure of the ring of quasimodular (or almost holomorphic modular) forms. Let $F = \sum f_r(-4\pi y)^{-r}$ and $f = f_0$ be an almost holomorphic modular form of weight k and depth $\le p$ and the quasimodular form which corresponds to it. One sees easily using the properties above that each coefficient f_r is quasimodular (of weight $k-2r$ and depth $\le p-r$) and that, if $\delta : \widetilde{M}_* \to \widetilde{M}_*$ is the map which sends f to f_1, then $f_r = \delta^r f/r!$ for all $r \ge 0$ (and $\delta^r f = 0$ for $r > p$), so that the expansion of $F(z)$ in powers of $-1/4\pi y$ is a kind of Taylor expansion formula. This gives us three operators from $\widetilde{M}_*$ to itself: the differentiation operator D, the operator E

which multiplies a quasimodular form of weight k by k, and the operator δ. Each of these operators is a derivation on the ring of quasimodular forms, and they satisfy the three commutation relations

$$[E,D] \,=\, 2\,D\,, \qquad [D,\delta] \,=\, -2\,\delta\,, \qquad [D,\delta] \,=\, E$$

(of which the first two just say that D and δ raise and lower the weight of a quasimodular form by 2, respectively), giving to $\widetilde{M}_*$ the structure of an $\mathfrak{sl}_2\mathbb{C})$-module. Of course the ring $\widehat{M}_*$, being isomorphic to $\widetilde{M}_*$, also becomes an $\mathfrak{sl}_2\mathbb{C})$-module, the corresponding operators being ϑ ($=\vartheta_k$ on $\widetilde{M}_k$), E (= multiplication by k on $\widehat{M}_k$) and δ^* (= "derivation with respect to $-1/4\pi y$"). From this point of view, the subspace M_* of $\widetilde{M}_*$ or $\widehat{M}_*$ appears simply as the kernel of the lowering operator δ (or δ^*). Using this $\mathfrak{sl}_2\mathbb{C})$-module structure makes many calculations with quasimodular or almost holomorphic modular forms simpler and more transparent.

♠ Counting Ramified Coverings of the Torns

We end this subsection by describing very briefly a beautiful and unexpected context in which quasimodular forms occur. Define

$$\Theta(X,z,\zeta) \;=\; \prod_{n>0}(1-q^n)\prod_{\substack{n>0\\ n\text{ odd}}}\bigl(1-e^{n^2X/8}\,q^{n/2}\,\zeta\bigr)\bigl(1-e^{-n^2X/8}\,q^{n/2}\,\zeta^{-1}\bigr)\,,$$

expand $\Theta(X,z,\zeta)$ as a Laurent series $\sum_{n\in\mathbb{Z}}\Theta_n(X,z)\,\zeta^n$, and expand $\Theta_0(X,z)$ as a Taylor series $\sum_{r=0}^{\infty}A_r(z)\,X^{2r}$. Then a somewhat intricate calculation involving Eisenstein series, theta series and quasimodular forms on $\Gamma(2)$ shows that each A_r is a quasimodular form of weight $6r$ on $\mathrm{SL}(2,\mathbb{Z})$, i.e., a weighted homogeneous polynomial in E_2, E_4, and E_6. In particular, $A_0(z)=1$ (this is a consequence of the Jacobi triple product identity, which says that $\Theta_n(0,z)=(-1)^nq^{n^2/2}$), so we can also expand $\log\Theta_0(X,z)$ as $\sum_{r=1}^{\infty}F_r(z)\,X^{2r}$ and $F_r(z)$ is again quasimodular of weight $6r$. These functions arose in the study of the "one-dimensional analogue of mirror symmetry": the coefficient of q^m in F_r counts the generically ramified coverings of degree m of a curve of genus 1 by a curve of genus $r+1$. ("Generic" means that each point has $\ge m-1$ preimages.) We thus obtain:

Theorem. *The generating function of generically ramified coverings of a torus by a surface of genus $g>1$ is a quasimodular form of weight $6g-6$ on $SL(2,\mathbb{Z})$.*

As an example, we have $F_1 = A_1 = \frac{1}{103680}(10E_2^3-6E_2E_4-4E_6) = q^2+8q^3+30q^4+80q^5+180q^6+\cdots$. In this case (but for no higher genus), the function $F_1=\frac{1}{1440}(E_4'+10E_2'')$ is also a linear combination of derivatives of Eisenstein series and we get a simple explicit formula $n(\sigma_3(n)-n\sigma_1(n))/6$ for the (correctly counted) number of degree n coverings of a torus by a surface of genus 2. ♡

5.4 Linear Differential Equations and Modular Forms

The statement of Proposition 16 is very simple, but not terribly useful, because it is difficult to derive properties of a function from a non-linear differential equation. We now prove a much more useful fact: if we express a modular form as a function, not of z, but of a modular function of z (i.e., a meromorphic modular form of weight zero), then it always satisfies a *linear* differentiable equation of finite order with algebraic coefficients. Of course, a modular form cannot be written as a single-valued function of a modular function, since the latter is invariant under modular transformations and the former transforms with a non-trivial automorphy factor, but in can be expressed locally as such a function, and the global non-uniqueness then simply corresponds to the monodromy of the differential equation.

The fact that we have mentioned is by no means new – it is at the heart of the original discovery of modular forms by Gauss and of the later work of Fricke and Klein and others, and appears in modern literature as the theory of Picard–Fuchs differential equations or of the Gauss–Manin connection – but it is not nearly as well known as it ought to be. Here is a precise statement.

Proposition 21. *Let $f(z)$ be a (holomorphic or meromorphic) modular form of positive weight k on some group Γ and $t(z)$ a modular function with respect to Γ. Express $f(z)$ (locally) as $\Phi(t(z))$. Then the function $\Phi(t)$ satisfies a linear differential equation of order $k+1$ with algebraic coefficients, or with polynomial coefficients if $\Gamma\backslash\mathfrak{H}$ has genus 0 and $t(z)$ generates the field of modular functions on Γ.*

This proposition is perhaps the single most important source of applications of modular forms in other branches of mathematics, so with no apology we sketch three different proofs, each one giving us different information about the differential equation in question.

Proof 1: We want to find a linear relation among the derivatives of f with respect to t. Since $f(z)$ is not defined in $\Gamma\backslash\mathfrak{H}$, we must replace d/dt by the operator $D_t = t'(z)^{-1}d/dz$, which makes sense in $\mathfrak{H}$. We wish to show that the functions $D_t^n f$ $(n = 0, 1, \ldots, k+1)$ are linearly dependent over the field of modular functions on Γ, since such functions are algebraic functions of $t(z)$ in general and rational functions in the special cases when $\Gamma\backslash\mathfrak{H}$ has genus 0 and $t(z)$ is a "Hauptmodul" (i.e., $t : \Gamma\backslash\mathfrak{H} \to \mathbb{P}_1(\mathbb{C})$ is an isomorphism). The difficulty is that, as seen in §5.1, differentiating the transformation equation (2) produces undesired extra terms as in (52). This is because the automorphy factor $(cz+d)^k$ in (2) is non-constant and hence contributes non-trivially when we differentiate. To get around this, we replace $f(z)$ by the *vector-valued* function $F : \mathfrak{H} \to \mathbb{C}^{k+1}$ whose mth component $(0 \le m \le k)$ is $F_m(z) = z^{k-m}f(z)$. Then

$$F_m\left(\frac{az+b}{cz+d}\right) \;=\; (az+b)^{k-m}\,(cz+d)^m\,f(z) \;=\; \sum_{n=0}^{m} M_{mn}\,F_n(z)\,, \qquad (71)$$

for all $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$, where $M = S^k(\gamma) \in \mathrm{SL}(k+1,\mathbb{Z})$ denotes the kth symmetric power of γ. In other words, we have $F(\gamma z) = MF(z)$ where the new "automorphy factor" M, although it is more complicated than the automorphy factor in (2) because it is a matrix rather than a scalar, is now independent of z, so that when we differentiate we get simply $(cz+d)^{-2}F'(\gamma z) = MF'(z)$. Of course, this equation contains $(cz+d)^{-2}$, so differentiating again with respect to z would again produce unwanted terms. But the Γ-invariance of t implies that $t'\bigl(\frac{az+b}{cz+d}\bigr) = (cz+d)^2 t'(z)$, so the factor $(cz+d)^{-2}$ is cancelled if we replace d/dz by $D_t = d/dt$. Thus (71) implies $D_tF(\gamma z) = MD_tF(z)$, and by induction $D_t^rF(\gamma z) = MD_t^rF(z)$ for all $r \ge 0$. Now consider the $(k+2)\times(k+2)$ matrix

$$\begin{pmatrix} f & D_tf & \cdots & D_t^{k+1}f \\ F & D_tF & \cdots & D_t^{k+1}F \end{pmatrix}.$$

The top and bottom rows of this matrix are identical, so its determinant is 0. Expanding by the top row, we find $0 = \sum_{n=0}^{k+1}(-1)^n \det(A_n(z))\, D_t^n f(z)$, where $A_n(z)$ is the $(k+1)\times(k+1)$ matrix $\bigl(F \;\, D_tF \;\, \cdots \;\, \widehat{D_t^nF} \;\, \cdots \;\, D_t^{k+1}F\bigr)$. From $D_t^rF(\gamma z) = MD_t^rF(z)$ $(\forall r)$ we get $A_n(\gamma z) = MA_n(z)$ and hence, since $\det(M) = 1$, that $A_n(z)$ is a Γ-invariant function. Since $A_n(z)$ is also meromorphic (including at the cusps), it is a modular function on Γ and hence an algebraic or rational function of $t(z)$, as desired. The advantage of this proof is that it gives us all $k+1$ linearly independent solutions of the differential equation satisfied by $f(z)$: they are simply the functions $f(z)$, $zf(z), \dots, z^kf(z)$.

Proof 2 (following a suggestion of Ouled Azaiez): We again use the differentiation operator D_t, but this time work with quasimodular rather than vector-valued modular forms. As in §5.2, choose a quasimodular form $\phi(z)$ of weight 2 on Γ with $\delta(\phi) = 1$ (e.g., $\frac{1}{12}E_2$ if $\Gamma = \Gamma_1$), and write each $D_t^nf(z)$ $(n = 0, 1, 2, \dots)$ as a polynomial in $\phi(z)$ with (meromorphic) modular coefficients. For instance, for $n = 1$ we find $D_tf \;=\; k\,\dfrac{f}{t'}\,\phi + \dfrac{\vartheta_\phi f}{t'}$ where ϑ_ϕ denotes the Serre derivative with respect to ϕ. Using that $\phi' - \phi^2 \in M_4$, one finds by induction that each D_t^nf is the sum of $k(k-1)\cdots(k-n+1)\,(\phi/t')^n f$ and a polynomial of degree $< n$ in ϕ. It follows that the $k+2$ functions $\bigl\{D_t^n(f)/f\bigr\}_{0\le n\le k+1}$ are linear combinations of the $k+1$ functions $\{(\phi/t')^n\}_{0\le n\le k}$ with modular functions as coefficients, and hence that they are linearly dependent over the field of modular functions on Γ.

Proof 3: The third proof will give an explicit differential equation satisfied by f. Consider first the case when f has weight 1. The funcion $t' = D(t)$ is a (meromorphic) modular form of weight 2, so we can form the Rankin–Cohen brackets $[f,t']_1$ and $[f,f]_2$ of weights 5 and 6, respectively, and the quotients $A = \frac{[f,t']_1}{ft'^2}$ and $B = -\frac{[f,f]_2}{2f^2t'^2}$ of weight 0. Then

$$D_t^2f \,+\, A\,D_tf \,+\, Bf \;=\; \frac{1}{t'}\left(\frac{f'}{t'}\right)' + \frac{ft''-2f't'}{ft'^2}\,\frac{f'}{t'} - \frac{ff''-2f'^2}{t'^2f^2}\,f \;=\; 0\,.$$

Since A and B are modular functions, they are rational (if t is a Hauptmodul) or algebraic (in any case) functions of t, say $A(z) = a(t(z))$ and $B(z) = b(t(z))$, and then the function $\varPhi(t)$ defined locally by $f(z) = \varPhi(t(z))$ satisfies $\varPhi''(t) + a(t)\varPhi'(t) + b(t)\varPhi(t) = 0$.

Now let f have arbitrary (integral) weight $k > 0$ and apply the above construction to the function $h = f^{1/k}$, which is formally a modular form of weight 1. Of course h is not really a modular form, since in general it is not even a well-defined function in the upper half-plane (it changes by a kth root of unity when we go around a zero of f). But it is defined locally, and the functions $A = \frac{[h,t']_1}{ht'^2}$ and $B = -\frac{[h,h]_2}{2h^2t'^2}$ *are* well-defined, because they are both homogeneous of degree 0 in h, so that the kth roots of unity occurring when we go around a zero of f cancel out. (In fact, a short calculation shows that they can be written directly in terms of Rankin–Cohen brackets of f and t', namely $A = \frac{[f,t']_1}{kft'^2}$ and $B = -\frac{[f,f]_2}{k^2(k+1)f^2t'^2}$.) Now just as before $A(z) = a(t(z))$, $B(z) = b(t(z))$ for some rational or algebraic functions $a(t)$ and $b(t)$. Then $\varPhi(t)^{1/k}$ is annihilated by the second order differential operator $L = d^2/dt^2 + a(t)d/dt + b(t)$ and $\varPhi(t)$ itself is annihilated by the $(k+1)$st order differential operator $\mathrm{Sym}^k(L)$ whose solutions are the kth powers or k-fold products of solutions of the differential equation $L\varPsi = 0$. The coefficients of this operator can be given by explicit expressions in terms of a and b and their derivatives (for instance, for $k = 2$ we find $\mathrm{Sym}^2(L) = d^3/dt^3 + 3a\,d^2/dt^2 + (a' + 2a^2 + 4b)\,d/dt + 2(b' + 2ab)$), and these in turn can be written as weight 0 quotients of appropriate Rankin–Cohen brackets.

Here are two classical examples of Proposition 21. For the first, we take $\varGamma = \varGamma(2)$, $f(z) = \theta_3(z)^2$ (with $\theta_3(z) = \sum q^{n^2/2}$ as in (32)) and $t(z) = \lambda(z)$, where $\lambda(z)$ is the Legendre modular function

$$\lambda(z) \;=\; 16\,\frac{\eta(z/2)^8\,\eta(2z)^{16}}{\eta(z)^{24}} \;=\; 1 \,-\, \frac{\eta(z/2)^{16}\,\eta(2z)^8}{\eta(z)^{24}} \;=\; \left(\frac{\theta_2(z)}{\theta_3(z)}\right)^4, \tag{72}$$

which is known to be a Hauptmodul for the group $\varGamma(2)$. Then

$$\theta_3(z)^2 \;=\; \sum_{n=0}^{\infty}\binom{2n}{n}\left(\frac{\lambda(z)}{16}\right)^n \;=\; F\left(\frac{1}{2},\frac{1}{2};\,1;\,\lambda(z)\right), \tag{73}$$

where $F(a,b;c;x) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{n!\,(c)_n}\,x^n$ with $(a)_n$ as in eq. (56) denotes the Gauss hypergeometric function, which satisfies the second order differential equation $x(x-1)\,y'' + \big((a+b+1)x - c\big)\,y' + aby = 0$. For the second example we take $\varGamma = \varGamma_1$, $t(z) = 1728/j(z)$, and $f(z) = E_4(z)$. Since f is a modular form of weight 4, it should satisfy a fifth order linear differential equation with respect to $t(z)$, but by the third proof above one should even have that the fourth root of f satisfies a second order differential equation, and indeed one finds

$$\sqrt[4]{E_4(z)} \;=\; F\bigl(\tfrac{1}{12},\,\tfrac{5}{12};\,1;\,t(z)\bigr) \;=\; 1+\frac{1\cdot 5}{1\cdot 1}\,\frac{12}{j(z)}+\frac{1\cdot 5\cdot 13\cdot 17}{1\cdot 1\cdot 2\cdot 2}\,\frac{12^2}{j(z)^2}+\cdots\,, \tag{74}$$

a classical identity which can be found in the works of Fricke and Klein.

♠ The Irrationality of $\zeta(3)$

In 1978, Roger Apéry created a sensation by proving:

Theorem. *The number* $\zeta(3)=\sum_{n\ge1} n^{-3}$ *is irrational.*

What he actually showed was that, if we define two sequences $\{a_n\} = \{1,\,5,\,73,\,1445,\,\dots\}$ and $\{b_n\}=\{0,\,6,\,351/4,\,62531/36,\,\dots\}$ as the solutions of the recursion

$$(n+1)^3\,u_{n+1} \;=\; (34n^3+51n^2+27n+5)\,u_n \;-\; n^3\,u_{n-1} \qquad (n\ge 1) \tag{75}$$

with initial conditions $a_0=1$, $a_1=5$, $b_0=0$, $b_1=6$, then we have

$$a_n\in\mathbb{Z}\quad(\forall n\ge 0)\,,\qquad D_n^3\,b_n\in\mathbb{Z}\quad(\forall n\ge 0)\,,\qquad \lim_{n\to\infty}\frac{b_n}{a_n} \;=\; \zeta(3)\,, \tag{76}$$

where D_n denotes the least common multiple of $1,2,\dots,n$. These three assertions together imply the theorem. Indeed, both a_n and b_n grow like C^n by the recursion, where $C=33.97\dots$ is the larger root of $C^2-34C+1=0$. On the other hand, $a_{n-1}b_n-a_nb_{n-1}=6/n^3$ by the recursion and induction, so the difference between b_n/a_n and its limiting value $\zeta(3)$ decreases like C^{-2n}. Hence the quantity $x_n=D_n^3(b_n-a_n\zeta(3))$ grows like by $D_n^3/(C+\mathrm{o}(1))^n$, which tends to 0 as $n\to\infty$ since $C>e^3$ and $D_n^3=(e^3+\mathrm{o}(1))^n$ by the prime number theorem. (Chebyshev's weaker elementary estimate of D_n would suffice here.) But if $\zeta(3)$ were rational then the first two statements in (76) would imply that the x_n are rational numbers with bounded denominators, and this is a contradiction.

Apéry's own proof of the three properties (76), which involved complicated explicit formulas for the numbers a_n and b_n as sums of binomial coefficients, was very ingenious but did not give any feeling for why any of these three properties hold. Subsequently, two more enlightening proofs were found by Frits Beukers, one using representations of a_n and b_n as multiple integrals involving Legendre polynomials and the other based on modularity. We give a brief sketch of the latter one. Let $\Gamma=\Gamma_0^+(6)$ as in §3.1 be the group $\Gamma_0(6)\cup\Gamma_0(6)W$, where $W=W_6=\frac{1}{\sqrt6}\left(\begin{smallmatrix}0&-1\\6&0\end{smallmatrix}\right)$. This group has genus 0 and the Hauptmodul

$$t(z) \;=\; \left(\frac{\eta(z)\,\eta(6z)}{\eta(2z)\,\eta(3z)}\right)^{12} \;=\; q\;-\;12\,q^2\;+\;66\,q^3\;-\;220\,q^4\;+\;\dots\,,$$

where $\eta(z)$ is the Dedekind eta-function defined in (34). For $f(z)$ we take the function

$$f(z) \;=\; \frac{\bigl(\eta(2z)\,\eta(3z)\bigr)^7}{\bigl(\eta(z)\,\eta(6z)\bigr)^5} \;=\; 1 \,+\, 5\,q \,+\, 13\,q^2 \,+\, 23\,q^3 \,+\, 29\,q^4 + \dots\,,$$

which is a modular form of weight 2 on Γ (and in fact an Eisenstein series, namely $5\mathbb{G}_2(z) - 2\mathbb{G}_2(2z) + 3\mathbb{G}_2(3z) - 30\mathbb{G}_2(6z)$). Proposition 21 then implies that if we expand $f(z)$ as a power series $1 + 5\,t(z) + 73\,t(z)^2 + 1445\,t(z)^3 + \cdots$ in $t(z)$, then the coefficients of the expansion satisfy a linear recursion with polynomial coefficients (this is equivalent to the statement that the power series itself satisfies a differential equation with polynomial coefficients), and if we go through the proof of the proposition to calculate the differential equation explicitly, we find that the recursion is exactly the one defining a_n. Hence $f(z) = \sum_{n=0}^{\infty} a_n\, t(z)^n$, and the integrality of the coefficients a_n follows immediately since $f(z)$ and $t(z)$ have integral coefficients and t has leading coefficient 1.

To get the properties of the second sequence $\{b_n\}$ is a little harder. Define

$$\begin{aligned} g(z) \;&=\; \mathbb{G}_4(z) - 28\,\mathbb{G}_4(2z) + 63\,\mathbb{G}_4(3z) - 36\,\mathbb{G}_4(6z) \\ &=\; q \,-\, 14\,q^2 \,+\, 91\,q^3 \,-\, 179\,q^4 \,+\, \cdots\,, \end{aligned}$$

an Eisenstein series of weight 4 on Γ. Write the Fourier expansion as $g(z) = \sum_{n=1}^{\infty} c_n\, q^n$ and define $\tilde{g}(z) = \sum_{n=1}^{\infty} n^{-3} c_n\, q^n$, so that $\tilde{g}''' = g$. This is the so-called Eichler integral associated to g and inherits certain modular properties from the modularity of g. (Specifically, the difference $\tilde{g}|_{-2}\,\gamma - \tilde{g}$ is a polynomial of degree ≤ 2 for every $\gamma \in \Gamma$, where $|_{-2}$ is the slash operator defined in (8).) Using this (we skip the details), one finds by an argument analogous to the proof of Proposition 21 that if we expand the product $f(z)\tilde{g}(z)$ as a power series $t(z) + \frac{117}{8}\,t(z)^2 + \frac{62531}{216}\,t(z)^3 + \cdots$ in $t(z)$, then this power series again satisfies a differential equation. This equation turns out to be the same one as the one satisfied by f (but with right-hand side 1 instead of 0), so the coefficients of the new expansion satisfy the same recursion and hence (since they begin 0, 1) must be one-sixth of Apéry's coefficients b_n, i.e., we have $6f(z)\tilde{g}(z) = \sum_{n=0}^{\infty} b_n\, t(z)^n$. The integrality of $D_n^3 b_n$ (indeed, even of $D_n^3 b_n/6$) follows immediately: the coefficients c_n are integral, so D_n^3 is a common denominator for the first n terms of $f\tilde{g}$ as a power series in q and hence also as a power series in $t(z)$. Finally, from the definition (13) of $\mathbb{G}_4(z)$ we find that the Fourier coefficients of g are given by the Dirichlet series identity

$$\sum_{n=1}^{\infty} \frac{c_n}{n^s} \;=\; \left(1 - \frac{28}{2^s} + \frac{63}{3^s} - \frac{36}{6^s}\right)\,\zeta(s)\,\zeta(s-3)\,,$$

so the limiting value of b_n/a_n (which must exist because $\{a_n\}$ and $\{b_n\}$ satisfy the same recursion) is given by

$$\frac{1}{6}\lim_{n\to\infty}\frac{b_n}{a_n} = \tilde{g}(z)\Big|_{t(z)=1/C} = \sum_{n=1}^{\infty}\frac{c_n}{n^3}\,q^n\Big|_{q=1} = \sum_{n=1}^{\infty}\frac{c_n}{n^s}\Big|_{s=3} = \frac{1}{6}\,\zeta(3)\,,$$

proving also the last assertion in (76). ♡

♠ An Example Coming from Percolation Theory

Imagine a rectangle R of width $r > 0$ and height 1 on which has been drawn a fine square grid (of size roughly $rN \times N$ for some integer N going to infinity). Each edge of the grid is colored black or white with probability 1/2 (the critical probability for this problem). The (horizontal) *crossing probability* $\Pi(r)$ is then defined as the limiting value, as $N \to \infty$, of the probability that there exists a path of black edges connecting the left and right sides of the rectangle. A simple combinatorial argument shows that there is always either a vertex-to-vertex path from left to right passing only through black edges or else a square-to-square path from top to bottom passing only through white edges, but never both. This implies that $\Pi(r) + \Pi(1/r) = 1$. The hypothesis of conformality which, though not proved, is universally believed and is at the basis of the modern theory of percolation, says that the corresponding problem, with R replaced by any (nice) open domain in the plane, is unchanged under conformal (biholomorphic) mappings, and this can be used to compute the crossing probability as the solution of a differential equation. The result, due to J. Cardy, is the formula $\Pi(r) = 2\pi\sqrt{3}\,\Gamma(\frac{1}{3})^{-3}\,t^{1/3}\,F\big(\frac{1}{3},\,\frac{2}{3};\,\frac{4}{3};\,t\big)$, where t is the cross-ratio of the images of the four vertices of the rectangle R when it is mapped biholomorphically onto the unit disk by the Riemann uniformization theorem. This cross-ratio is known to be given by $t = \lambda(ir)$ with $\lambda(z)$ as in (72). In modular terms, using Proposition 21, we find that this translates to the formula $\Pi(r) = -2^{7/3}\,3^{-1/2}\,\pi^2\,\Gamma(\frac{1}{3})^{-3}\int_r^{\infty}\eta(iy)^4dy\,$; i.e., the derivative of $\Pi(r)$ is essentially the restriction to the imaginary axis of the modular form $\eta(z)^4$ of weight 2. Conversely, an easy argument using modular forms on $\mathrm{SL}(2,\mathbb{Z})$ shows that Cardy's function is the *unique* function satisfying the functional equation $\Pi(r) + \Pi(1/r) = 1$ and having an expansion of the form $e^{-2\pi\alpha r}$ times a power series in $e^{-2\pi r}$ for some $\alpha \in \mathbb{R}$ (which is then given by $\alpha = 1/6$). Unfortunately, there seems to be no physical argument implying *a priori* that the crossing probability has the latter property, so one cannot (yet?) use this very simple modular characterization to obtain a new proof of Cardy's famous formula. ♡

6 Singular Moduli and Complex Multiplication

The theory of complex multiplication, the last topic which we will treat in detail in these notes, is by any standards one of the most beautiful chapters in all of number theory. To describe it fully one needs to combine themes relating to elliptic curves, modular forms, and algebraic number theory. Given the emphasis of these notes, we will discuss mostly the modular forms side, but in this introduction we briefly explain the notion of complex multiplication in the language of elliptic curves.

An elliptic curve over $\mathbb{C}$, as discussed in §1, can be represented by a quotient $E = \mathbb{C}/\Lambda$, where Λ is a lattice in $\mathbb{C}$. If $E' = \mathbb{C}/\Lambda'$ is another curve and λ a complex number with $\lambda\Lambda \subseteq \Lambda'$, then multiplication by λ induces an algebraic map from E to E'. In particular, if $\lambda\Lambda \subseteq \Lambda$, then we get a map from E to itself. Of course, we can always achieve this with $\lambda \in \mathbb{Z}$, since Λ is a $\mathbb{Z}$-module. These are the only possible real values of λ, and for generic lattices also the only possible complex values. Elliptic curves $E = \mathbb{C}/\Lambda$ where $\lambda\Lambda \subseteq \Lambda$ for some non-real value of λ are said to *admit complex multiplication.*

As we have seen in these notes, there are two completely different ways in which elliptic curves are related to modular forms. On the one hand, the moduli space of elliptic curves is precisely the domain of definition $\Gamma_1\backslash\mathfrak{H}$ of modular functions on the full modular group, via the map $\Gamma_1 z \leftrightarrow \bigl[\mathbb{C}/\Lambda_z\bigr]$, where $\Lambda_z = \mathbb{Z}z + \mathbb{Z} \subset \mathbb{C}$. On the other hand, elliptic curves over $\mathbb{Q}$ are supposed (and now finally known) to have parametrizations by modular functions and to have Hasse-Weil L-functions that coincide with the Hecke L-series of certain cusp forms of weight 2. The elliptic curves with complex multiplication are of special interest from both points of view. If we think of $\mathfrak{H}$ as parametrizing elliptic curves, then the points in $\mathfrak{H}$ corresponding to elliptic curves with complex multiplication (usually called *CM points* for short) are simply the numbers $\mathfrak{z} \in \mathfrak{H}$ which satisfy a quadratic equation over $\mathbb{Z}$. The basic fact is that the value of $j(\mathfrak{z})$ (or of any other modular function with algebraic coefficients evaluated at $\mathfrak{z}$) is then an algebraic number; this says that an elliptic curve with complex multiplication is always defined over $\overline{\mathbb{Q}}$ (i.e., has a Weierstrass equation with algebraic coefficients). Moreover, these special algebraic numbers $j(\mathfrak{z})$, classically called *singular moduli*, have remarkable properties: they give explicit generators of the class fields of imaginary quadratic fields, the differences between them factor into small prime factors, their traces are themselves the coefficients of modular forms, etc. This will be the theme of the first two subsections. If on the other hand we consider the L-function of a CM elliptic curve and the associated cusp form, then again both have very special properties: the former belongs to two important classes of number-theoretical Dirichlet series (Epstein zeta functions and L-series of grossencharacters) and the latter is a theta series with spherical coefficients associated to a binary quadratic form. This will lead to several applications which are treated in the final two subsections.

6.1 Algebraicity of Singular Moduli

In this subsection we will discuss the proof, refinements, and applications of the following basic statement:

Proposition 22. *Let $\mathfrak{z} \in \mathfrak{H}$ be a CM point. Then $j(\mathfrak{z})$ is an algebraic number.*

Proof. By definition, $\mathfrak{z}$ satisfies a quadratic equation over $\mathbb{Z}$, say $A\mathfrak{z}^2 + B\mathfrak{z} + C = 0$. There is then always a matrix $M \in M(2,\mathbb{Z})$, not proportional to the identity, which fixes $\mathfrak{z}$. (For instance, we can take $M = \left(\begin{smallmatrix} B & C \\ -A & 0 \end{smallmatrix}\right)$.) This matrix has a positive determinant, so it acts on the upper half-plane in the usual way. The two functions $j(z)$ and $j(Mz)$ are both modular functions on the subgroup $\Gamma_1 \cap M^{-1}\Gamma_1 M$ of finite index in Γ_1, so they are algebraically dependent, i.e., there is a non-zero polynomial $P(X,Y)$ in two variables such that $P(j(Mz), j(z))$ vanishes identically. By looking at the Fourier expansion at ∞ we can see that the polynomial P can be chosen to have coefficients in $\mathbb{Q}$. (We omit the details, since we will prove a more precise statement below.) We can also assume that the polynomial $P(X,X)$ is not identically zero. (If some power of $X-Y$ divides P then we can remove it without affecting the validity of the relation $P(j(Mz), j(z)) \equiv 0$, since $j(Mz)$ is not identically equal to $j(z)$.) The number $j(\mathfrak{z})$ is then a root of the non-trivial polynomial $P(X,X) \in \mathbb{Q}[X]$, so it is an algebraic number.

More generally, if $f(z)$ is any modular function (say, with respect to a subgroup of finite index of $\mathrm{SL}(2,\mathbb{Z})$) with algebraic Fourier coefficients in its q-expansion at infinity, then $f(\mathfrak{z}) \in \overline{\mathbb{Q}}$, as one can see either by showing that $f(Mz)$ and $f(z)$ are algebraically dependent over $\overline{\mathbb{Q}}$ for any $M \in M(2,\mathbb{Z})$ with $\det M > 0$ or by showing that $f(z)$ and $j(z)$ are algebraically dependent over $\overline{\mathbb{Q}}$ and using Proposition 22. The full theory of complex multiplication describes precisely the number field in which these numbers $f(\mathfrak{z})$ lie and the way that $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ acts on them. (Roughly speaking, any Galois conjugate of any $f(\mathfrak{z})$ has the form $f^*(\mathfrak{z}^*)$ for some other modular form with algebraic coefficients and CM point $\mathfrak{z}^*$, and there is a recipe to compute both.) We will not explain anything about this in these notes, except for a few words in the third application below, but refer the reader to the texts listed in the references.

We now return to the j-function and the proof above. The key point was the algebraic relation between $j(z)$ and $j(Mz)$, where M was a matrix in $M(2,\mathbb{Z})$ of positive determinant m fixing the point $\mathfrak{z}$. We claim that the polynomial P relating $j(Mz)$ and $j(z)$ can be chosen to depend only on m. More precisely, we have:

Proposition 23. *For each $m \in \mathbb{N}$ there is a polynomial $\Psi_m(X,Y) \in \mathbb{Z}[X,Y]$, symmetric up to sign in its two arguments and of degree $\sigma_1(m)$ with respect to either one, such that $\Psi_m(j(Mz), j(z)) \equiv 0$ for every matrix $M \in M(2,\mathbb{Z})$ of determinant m.*

Proof. Denote by $\mathcal{M}_m$ the set of matrices in $M(2,\mathbb{Z})$ of determinant m. The group Γ_1 acts on $\mathcal{M}_m$ by right and left multiplication, with finitely many orbits. More precisely, an easy and standard argument shows that the finite set

$$\mathcal{M}_m^* = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \mid a,\, b,\, d \in \mathbb{Z},\ ad = m,\ 0 \le b < d \right\} \subset \mathcal{M}_m \qquad (77)$$

is a full set of representatives for $\varGamma_1\backslash\mathcal{M}_m$; in particular, we have

$$\bigl|\varGamma_1\backslash\mathcal{M}_m\bigr| \;=\; \bigl|\mathcal{M}_m^*\bigr| \;=\; \sum_{ad=m} d \;=\; \sigma_1(m)\,. \tag{78}$$

We claim that we have an identity

$$\prod_{M\in\varGamma_1\backslash\mathcal{M}_m} \bigl(X \,-\, j(Mz)\bigr) \;=\; \varPsi_m(X, j(z)) \qquad (z\in\mathfrak{H},\ X\in\mathbb{C}) \tag{79}$$

for some polynomial $\varPsi_m(X,Y)$. Indeed, the left-hand side of (79) is well-defined because $j(Mz)$ depends only on the class of M in $\varGamma_1\backslash\mathcal{M}_m$, and it is $\varGamma_1$-invariant because $\mathcal{M}_m$ is invariant under right multiplication by elements of $\varGamma_1$. Furthermore, it is a polynomial in X (of degree $\sigma_1(m)$, by (78)) each of whose coefficients is a holomorphic function of z and of at most exponential growth at infinity, since each $j(Mz)$ has these properties and each coefficient of the product is a polynomial in the $j(Mz)$. But a $\varGamma_1$-invariant holomorphic function in the upper half-plane with at most exponential growth at infinity is a polynomial in $j(z)$, so that we indeed have (79) for some polynomial $\varPsi_m(X,Y)\in\mathbb{C}[X,Y]$. To see that the coefficients of $\varPsi_m$ are in $\mathbb{Z}$, we use the set of representatives (77) and the Fourier expansion of j, which has the form $j(z) = \sum_{n=-1}^{\infty} c_n q^n$ with $c_n\in\mathbb{Z}$ ($c_{-1}=1$, $c_0=744$, $c_1=196884$ etc.). Thus

$$\begin{aligned}\varPsi_m(X,j(z)) \;&=\; \prod_{\substack{ad=m\\ d>0}}\prod_{b=0}^{d-1}\Bigl(X \,-\, j\Bigl(\frac{az+b}{d}\Bigr)\Bigr)\\ &=\; \prod_{\substack{ad=m\\ d>0}}\;\prod_{b \pmod d}\Bigl(X \,-\, \sum_{n=-1}^{\infty} c_n\,\zeta_d^{bn}\,q^{an/d}\Bigr),\end{aligned}$$

where q^α for $\alpha\in\mathbb{Q}$ denotes $e^{2\pi i\alpha z}$ and $\zeta_d = e^{2\pi i/d}$. The expression in parentheses belongs to the ring $\mathbb{Z}[\zeta_d][X][q^{-1/d},q^{1/d}]]$ of Laurent series in $q^{1/d}$ with coefficients in $\mathbb{Z}[\zeta_d]$, but applying a Galois conjugation $\zeta_d\mapsto\zeta_d^r$ with $r\in(\mathbb{Z}/d\mathbb{Z})^*$ just replaces b in the inner product by br, which runs over the same set $\mathbb{Z}/d\mathbb{Z}$, so the inner product has coefficients in $\mathbb{Z}$. The fractional powers of q go away at the same time (because the product over b is invariant under $z\mapsto z+1$), so each inner product, and hence $\varPsi_m(X,j(z))$ itself, belongs to $\mathbb{Z}[X][q^{-1},q]]$. Now the fact that it is a polynomial in $j(z)$ and that $j(z)$ has a Fourier expansion with integral coefficients and leading coefficient q^{-1} imlies that $\varPsi_m(X,j(z))\in\mathbb{Z}[X,j(z)]$. Finally, the symmetry of $\varPsi_m(X,Y)$ up to sign follows because $z' = Mz$ with $M = \left(\begin{smallmatrix} a & b\\ c & d\end{smallmatrix}\right)\in\mathcal{M}_m$ is equivalent to $z = M'z'$ with $M' = \left(\begin{smallmatrix} d & -b\\ -c & a\end{smallmatrix}\right)\in\mathcal{M}_m$.

An example will make all of this clearer. For $m = 2$ we have

$$\prod_{M\in\Gamma_1\backslash\mathcal{M}_m} (X - j(z)) = \left(X - j\left(\frac{z}{2}\right)\right)\left(X - j\left(\frac{z+1}{2}\right)\right)(X - j(2z))$$

by (77). Write this as $X^3 - A(z)X^2 + B(z)X - C(z)$. Then

$$\begin{aligned} A(z) &= j\left(\frac{z}{2}\right) + j\left(\frac{z+1}{2}\right) + j(2z) \\ &= \left(q^{-1/2} + 744 + \mathrm{o}(q)\right) + \left(-q^{-1/2} + 744 + \mathrm{o}(q)\right) + \left(q^{-2} + 744 + \mathrm{o}(q)\right) \\ &= q^{-2} + 0\,q^{-1} + 2232 + \mathrm{o}(1) \\ &= j(z)^2 - 1488\,j(z) + 16200 + \mathrm{o}(1) \end{aligned}$$

as $z \to i\infty$, and since $A(z)$ is holomorphic and Γ_1-invariant this implies that $A = j^2 - 1488j + 16200$. A similar calculation gives $B = 1488j^2 + 40773375j + 8748000000$ and $C = -j^3 + 162000j^2 - 8748000000j + 157464000000000$, so

$$\begin{aligned} \Psi_2(X,Y) = &- X^2Y^2 + X^3 + 1488X^2Y + 1488XY^2 + Y^3 - 162000X^2 \\ &+ 40773375XY - 162000Y^2 + 8748000000X + 8748000000Y \\ &- 157464000000000\,. \end{aligned} \tag{80}$$

Remark. The polynomial $\Psi_m(X,Y)$ is not in general irreducible: if m is not square-free, then it factors into the product of all $\Phi_{m/r^2}(X,Y)$ with $r \in \mathbb{N}$ such that $r^2|m$, where $\Phi_m(X,Y)$ is defined exactly like $\Psi_m(X,Y)$ but with $\mathcal{M}_m$ replaced by the set $\mathcal{M}_m^0$ of primitive matrices of determinant m. The polynomial $\Phi_m(X,Y)$ is always irreducible.

To obtain the algebraicity of $j(\mathfrak{z})$, we used that this value was a root of $P(X,X)$. We therefore should look at the restriction of the polynomial $\Psi_m(X,Y)$ to the diagonal $X = Y$. In the example (80) just considered, two properties of this restriction are noteworthy. First, it is (up to sign) monic, of degree 4. Second, it has a striking factorization:

$$\Psi_2(X,X) = -(X - 8000)\cdot(X + 3375)^2\cdot(X - 1728)\,. \tag{81}$$

We consider each of these properties in turn. We assume that m is not a square, since otherwise $\Psi_m(X,X)$ is identically zero because $\Psi_m(X,Y)$ contains the factor $\Psi_1(X,Y) = X - Y$.

Proposition 24. *For m not a perfect square, the polynomial $\Psi_m(X,X)$ is, up to sign, monic of degree $\sigma_1^+(m) := \sum_{d|m} \max(d, m/d)$.*

Proof. Using the identity $\prod_{b \pmod d}(x - \zeta_d^b y) = x^d - y^d$ we find

$$\begin{aligned}\Psi_m(j(z), j(z)) &= \prod_{ad=m} \prod_{b \pmod d} \left(j(z) - j\left(\frac{az+b}{d}\right)\right) \\ &= \prod_{ad=m} \prod_{b \pmod d} \left(q^{-1} - \zeta_d^{-b}\, q^{-a/d} + \mathrm{o}(1)\right) \\ &= \prod_{ad=m} \left(q^{-d} - q^{-a} + \text{(lower order terms)}\right) \;\sim\; \pm\, q^{-\sigma_1^+(m)}\end{aligned}$$

as $\Im(z) \to \infty$, and this proves the proposition since $j(z) \sim q^{-1}$.

Corollary. *Singular moduli are algebraic integers.* □

Now we consider the factors of $\Psi_m(X, X)$. First let us identify the three roots in the factorization for $m = 2$ just given. The three CM points i, $(1 + i\sqrt{7})/2$ and $i\sqrt{2}$ are fixed, respectively, by the three matrices $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & -1 \\ 2 & 0 \end{smallmatrix}\right)$ of determinant 2, so each of the corresponding j-values must be a root of the polynomial (81). Computing these values numerically to low accuracy to see which is which, we find

$$j(i) \;=\; 1728\,, \qquad j\Bigl(\frac{1+i\sqrt{7}}{2}\Bigr) \;=\; -3375\,, \qquad j(i\sqrt{2}) \;=\; 8000\,.$$

(Another way to distinguish the roots, without doing any transcendental calculations, would be to observe, for instance, that i and $i\sqrt{2}$ are fixed by matrices of determinant 3 but $(1 + i\sqrt{7})/2$ is not, and that $X + 3375$ does not occur as a factor of $\Psi_3(X, X)$ but both $X - 1728$ and $X - 8000$ do.)

The same method can be used for any other CM point. Here is a table of the first few values of $j(\mathfrak{z}_D)$, where $\mathfrak{z}_D$ equals $\frac{1}{2}\sqrt{D}$ for D even and $\frac{1}{2}(1+\sqrt{D})$ for D odd:

D	-3	-4	-7	-8	-11	-12	-15	-16	-19
$j(\mathfrak{z}_D)$	0	1728	-3375	8000	-32768	54000	$-\frac{191025+85995\sqrt{5}}{2}$	287496	-884736

We can make this more precise. For each discriminant (integer congruent to 0 or 1 mod 4) $D < 0$ we consider, as at the end of §1.2, the set $\mathfrak{Q}_D$ of primitive positive definite binary quadratic forms of discriminant D, i.e., functions $Q(x, y) = Ax^2 + Bxy + Cy^2$ with A, B, $C \in \mathbb{Z}$, $A > 0$, $\gcd(A, B, C) = 1$ and $B^2 - 4AC = D$. To each $Q \in \mathfrak{Q}_D$ we associate the root $\mathfrak{z}_Q$ of $Q(\mathfrak{z}, 1) = 0$ in $\mathfrak{H}$. This gives a Γ_1-equivariant bijection between $\mathfrak{Q}_D$ and the set $\mathfrak{Z}_D \subset \mathfrak{H}$ of CM points of discriminant D. (The discriminant of a CM point is the smallest discriminant of a quadratic polynomial over $\mathbb{Z}$ of which it is a root.) In particular, the cardinality of $\Gamma_1\backslash\mathfrak{Z}_D$ is $h(D)$, the class number of D. We choose a set of representatives $\{\mathfrak{z}_{D,i}\}_{1\le i\le h(D)}$ for $\Gamma_1\backslash\mathfrak{Z}_D$ (e.g., the points of

$\mathfrak{Z}_D \cap \widetilde{\mathcal{F}_1}$, where $\widetilde{\mathcal{F}_1}$ is the fundamental domain constructed in §1.2, corresponding to the set $\mathfrak{Q}_D^{\mathrm{red}}$ in (5)), with $\mathfrak{z}_{D,1} = \mathfrak{z}_D$. We now form the *class polynomial*

$$H_D(X) \;=\; \prod_{\mathfrak{z}\in\Gamma_1\backslash\mathfrak{Z}_D} \bigl(X \,-\, j(\mathfrak{z})\bigr) \;=\; \prod_{1\le i\le h(D)} \bigl(X \,-\, j(\mathfrak{z}_{D,i})\bigr)\,. \tag{82}$$

Proposition 25. *The polynomial $H_D(X)$ belongs to $\mathbb{Z}[X]$ and is irreducible. In particular, the number $j(\mathfrak{z}_D)$ is algebraic of degree exactly $h(D)$ over $\mathbb{Q}$, with conjugates $j(\mathfrak{z}_{D,i})$ $(1 \le i \le h(D))$.*

Proof. We indicate only the main ideas of the proof. We have already proved that $j(\mathfrak{z})$ for any CM point $\mathfrak{z}$ is a root of the equation $\Psi_m(X,X) = 0$ whenever m is the determinant of a matrix $M \in M(2,\mathbb{Z})$ fixing $\mathfrak{z}$. The main point is that the set of these m's depends only on the discriminant D of $\mathfrak{z}$, not on $\mathfrak{z}$ itself, so that the different numbers $j(\mathfrak{z}_{D,i})$ are roots of the same equations and hence are conjugates. Let $A\mathfrak{z}^2 + B\mathfrak{z} + C = 0$ $(A > 0)$ be the minimal equation of $\mathfrak{z}$ and suppose that $M = \left(\begin{smallmatrix} a & b\\ c & d\end{smallmatrix}\right) \in \mathcal{M}_m$ fixes $\mathfrak{z}$. Then since $cz^2 + (d-a)z - b = 0$ we must have $(c, d-a, -b) = u(A,B,C)$ for some $u \in \mathbb{Z}$. This gives

$$M \;=\; \begin{pmatrix} \frac12(t-Bu) & -Cu\\ Au & \frac12(t+Bu)\end{pmatrix}\,, \qquad \det M \;=\; \frac{t^2 - Du^2}{4}\,, \tag{83}$$

where $t = \operatorname{tr} M$. Convesely, if t and u are any integers with $t^2 - Du^2 = 4m$, then (83) gives a matrix $M \in \mathcal{M}_m$ fixing $\mathfrak{z}$. Thus the set of integers $m = \det M$ with $M\mathfrak{z} = \mathfrak{z}$ is the set of numbers $\frac14(t^2 - Du^2)$ with $t \equiv Du \pmod 2$ or, more invariantly, the set of norms of elements of the quadratic order

$$\mathcal{O}_D \;=\; \mathbb{Z}[\mathfrak{z}_D] \;=\; \left\{\frac{t+u\sqrt{D}}{2}\;\middle|\; t,\,u \in \mathbb{Z},\quad t \equiv Du \pmod 2\right\} \tag{84}$$

of discriminant D, and this indeed depends only on D, not on $\mathfrak{z}$. We can then obtain $H_D(X)$, or at least its square, as the g.c.d. of finitely many polymials $\Psi_m(X,X)$, just as we obtained $H_{-7}(X)^2 = (X + 3375)^2$ as the g.c.d. of $\Psi_2(X,X)$ and $\Psi_3(X,X)$ in the example above. (Start, for example, with a prime m_1 which is the norm of an element of $\mathcal{O}_D$ – there are known to be infinitely many – and then choose finitely many further m's which are also norms of elements in $\mathcal{O}_D$ but not any of the finitely many other quadratic orders in which m_1 is a norm.)

This argument only shows that $H_D(X)$ has rational coefficients, not that it is irreducible. The latter fact is proved most naturally by studying the arithmetic of the corresponding elliptic curves with complex multiplication (roughly speaking, the condition of having complex multiplication by a given order $\mathcal{O}_D$ is purely algebraic and hence is preserved by Galois conjugation), but since the emphasis in these notes is on modular methods and their applications, we omit the details.

The proof just given actually yields the formula

$$\Psi_m(X,X) \;=\; \pm \prod_{D<0} H_D(X)^{r_D(m)/w(D)} \qquad (m \in \mathbb{N},\ m \neq \text{square})\,, \tag{85}$$

due to Kronecker. Here $r_D(m) = \bigl|\{\,(t,u) \in \mathbb{Z}^2 \mid t^2 - Du^2 = 4m\,\}\bigr| = \bigl|\{\lambda \in \mathcal{O}_D \mid N(\lambda) = m\}\bigr|$ and $w(D)$ is the number of units in $\mathcal{O}_D$, which is equal to 6 or 4 for $D = -3$ or -4, respectively, and to 2 otherwise. The product in (85) is finite since $r_D(m) \neq 0 \Rightarrow 4m = t^2 - u^2D \geq |D|$ because u can't be equal to 0 if m is not a square.

There is another form of formula (85) which will be used in the second application below. As well as the usual class number $h(D)$, one has the *Hurwitz class number* $h^*(D)$ (the traditional notation is $H(|D|)$), defined as the number of *all* Γ_1-equivalence classes of positive definite binary quadratic forms of discriminant D, not just the primitive ones, counted with multiplicity equal to one over the order of their stabilizer in Γ_1 (which is 2 or 3 if the corresponding point in the fundamental domain for $\Gamma_1\backslash\mathfrak{H}$ is at i or ρ and is 1 otherwise). In formulas, $h^*(D) = \sum_{r^2|D} h'(D/r^2)$, where the sum is over all $r \in \mathbb{N}$ for which $r^2|D$ (and for which D/r^2 is still congruent to 0 or 1 mod 4, since otherwise $h'(D/r^2)$ will be 0) and $h'(D) = h(D)/\frac{1}{2}w(D)$ with $w(D)$ as above. Similarly, we can define a modified class "polynomial" $H_D^*(X)$, of "degree" $h^*(D)$, by

$$H_D^*(X) \;=\; \prod_{r^2|D} H_{D/r^2}(X)^{2/w(D)}\,,$$

e.g., $H_{-12}^*(X) = X^{1/3}(X - 54000)$. (These are actual polynomials unless $|D|$ or $3|D|$ is a square.) Then (85) can be written in the following considerably simpler form:

$$\Psi_m(X,X) \;=\; \pm \prod_{t^2<4m} H_{t^2-4m}^*(X) \qquad (m \in \mathbb{N},\ m \neq \text{square})\,. \tag{86}$$

This completes our long discussion of the algebraicity of singular moduli. We now describe some of the many applications.

♠ Strange Approximations to π

We start with an application that is more fun than serious. The discriminant $D = -163$ has class number one (and is in fact known to be the smallest such discriminant), so Proposition 25 implies that $j(\mathfrak{z}_D)$ is a rational integer. Moreover, it is large (in absolute value) because $j(z) \approx q^{-1}$ and the $q = e^{2\pi i z}$ corresponding to $z = \mathfrak{z}_{163}$ is roughly -4×10^{-18}. But then from $j(z) = q^{-1} + 744 + \mathrm{O}(q)$ we find that q^{-1} is extremely close to an integer, giving the formula

$$e^{\pi\sqrt{163}} \;=\; 262537412640768743.99999999999250072597\cdots$$

which would be very startling if one did not know about complex multiplication. By taking logarithms, one gets an extremely good approximation to π:

$$\pi \;=\; \frac{1}{\sqrt{163}}\,\log(262537412640768744) \;-\; \bigl(2.237\cdots\times 10^{-31}\bigr)\,.$$

(A poorer but simpler approximation, based on the fact that $j(\mathfrak{z}_{163})$ is a perfect cube, is $\pi \approx \frac{3}{\sqrt{163}}\log(640320)$, with an error of about 10^{-16}.) Many further identities of this type were found, still using the theory of complex multiplication, by Ramanujan and later by Dan Shanks, the most spectactular one being

$$\pi \;-\; \frac{6}{\sqrt{3502}}\log\bigl[2\,\varepsilon(x_1)\,\varepsilon(x_2)\,\varepsilon(x_3)\,\varepsilon(x_4)\bigr] \;\approx\; 7.4\times 10^{-82}$$

where $x_1 = 429+304\sqrt{2}$, $x_2 = \frac{627}{2}+221\sqrt{2}$, $x_3 = \frac{1071}{2}+92\sqrt{34}$, $x_4 = \frac{1553}{2}+133\sqrt{34}$, and $\varepsilon(x) = x+\sqrt{x^2-1}$. Of course these formulas are more curiosities than useful ways to compute π, since the logarithms are no easier to compute numerically than π is and in any case, if we allow complex numbers, then Euler's formula $\pi = \frac{1}{\sqrt{-1}}\log(-1)$ is an exact formula of the same kind! ♡

♠ Computing Class Numbers

By comparing the degrees on both sides of (85) or (86) and using Proposition 24, we obtain the famous *Hurwitz-Kronecker class number relations*

$$\sigma_1^+(m) \;=\; \sum_{D<0}\frac{h(D)}{w(D)}\,r_D(m) \;=\; \sum_{t^2<4m} h^*(t^2-4m) \qquad (m\in\mathbb{N},\ m\neq\text{square})\,. \tag{87}$$

(In fact the equality of the first and last terms is true also for m square if we replace the summation condition by $t^2\le 4m$ and define $h^*(0) = -\frac{1}{12}$, as one shows by a small modification of the proof given here.) We mention that this formula has a geometric interpretation in terms of intersection numbers. Let X denote the modular curve $\Gamma_1\backslash\mathfrak{H}$. For each $m\ge 1$ there is a curve $T_m\subset X\times X$, the *Hecke correspondence*, corresponding to the mth Hecke operator T_m introduced in §4.1. (The preimage of this curve in $\mathfrak{H}\times\mathfrak{H}$ consists of all pairs (z, Mz) with $z\in\mathfrak{H}$ and $M\in\mathcal{M}_m$.) For $m=1$, this curve is just the diagonal. Now the middle or right-hand term of (87) counts the "physical" intersection points of T_m and T_1 in $X\times X$ (with appropriate multiplicities if the intersections are not transversal), while the left-hand term computes the same number homologically, by first compactifying X to $\bar{X} = X\cup\{\infty\}$ (which is isomorphic to $\mathbb{P}^1(\mathbb{C})$ via $z\mapsto j(z)$) and T_m and T_1 to their closures $\bar{T}_m$ and $\bar{T}_1$ in $\bar{X}\times\bar{X}$ and then computing the intersection number of the homology classes of $\bar{T}_m$ and $\bar{T}_1$ in $H_2(\bar{X}\times\bar{X})\cong\mathbb{Z}^2$ and correcting this by the contribution coming from intersections at infinity of the compactified curves.

Equation (87) gives a formula for $h^*(-4m)$ in terms of $h^*(D)$ with $|D| < 4m$. This does not quite suffice to calculate class numbers recursively since only half of all discriminants are multiples of 4. But by a quite similar type of argument (cf. the discussion in §6.2 below) one can prove a second class number relation, namely

$$\sum_{t^2 \le 4m} (m - t^2)\, h^*(t^2 - 4m) \;=\; \sum_{d|m} \min(d, m/d)^3 \qquad (m \in \mathbb{N}) \qquad (88)$$

(again with the convention that $h^*(0) = -\frac{1}{12}$), and this together with (87) *does* suffice to determine $h^*(D)$ for all D recursively, since together they express $h^*(-4m) + 2h^*(1 - 4m)$ and $mh^*(1 - 4m) + 2(m - 1)h^*(1 - 4m)$ as linear combinations of $h^*(D)$ with $|D| < 4m - 1$, and every negative discriminant has the form $-4m$ or $1 - 4m$ for some $m \in \mathbb{N}$. This method is quite reasonable computationally if one wants to compute a table of class numbers $h^*(D)$ for $-X < D < 0$, with about the same running time (viz., $\mathrm{O}(X^{3/2})$ operations) as the more direct method of counting all reduced quadratic forms with discriminant in this range. ♡

♠ Explicit Class Field Theory for Imaginary Quadratic Fields

Class field theory, which is the pinnacle of classical algebraic number theory, gives a complete classification of all abelian extensions of a given number field K. In particular, it says that the unramified abelian extensions (we omit all definitions) are the subfields of a certain finite extension H/K, the *Hilbert class field*, whose degree over K is equal to the class number of K and whose Galois group over K is canonically isomorphic to the class group of K, while the ramified abelian extensions have a similar description in terms of more complicated partititons of the ideals of $\mathcal{O}_K$ into finitely many classes. However, this theory, beautiful though it is, gives no method to actually *construct* the abelian extensions, and in fact it is only known how to do this explicitly in two cases: $\mathbb{Q}$ and imaginary quadratic fields. If $K = \mathbb{Q}$ then the Hilbert class field is trivial, since the class number is 1, and the ramified abelian extensions are just the subfields of $\mathbb{Q}(e^{2\pi i/N})$ $(N \in \mathbb{N})$ by the Kronecker-Weber theorem. For imaginary quadratic fields, the result (in the unramified case) is as follows.

Theorem. *Let K be an imaginary quadratic field, with discriminant D and Hilbert class field H. Then the $h(D)$ singular moduli $j(\mathfrak{z}_{D,i})$ are conjugate to one another over K (not just over $\mathbb{Q}$), any one of them generates H over K, and the Galois group of H over K permutes them transitively. More precisely, if we label the CM points of discriminant D by the ideal classes of K and identify* $\mathrm{Gal}(H/K)$ *with the class group of K by the fundamental isomorphism of class field theory, then for any two ideal classes $\mathcal{A}$, $\mathcal{B}$ of K the element $\sigma_{\mathcal{A}}$ of* $\mathrm{Gal}(H/K)$ *sends $j(\mathfrak{z}_{\mathcal{B}})$ to $j(\mathfrak{z}_{\mathcal{A}\mathcal{B}})$.*

The ramified abelian extensions can also be described completely by complex multiplication theory, but then one has to use the values of other modular functions (not just of $j(z)$) evaluated at all points of $K \cap \mathfrak{H}$ (not just those of discriminant D). Actually, even for the Hilbert class fields it can be advantageous to use modular functions other than $j(z)$. For instance, for $D = -23$, the first discriminant with class number not a power of 2 (and therefore the first non-trivial example, since Gauss's theory of genera describes the Hilbert class field of D when the class group has exponent 2 as a composite quadratic extension, e.g., H for $K = \mathbb{Q}(\sqrt{-15})$ is $\mathbb{Q}(\sqrt{-3},\,\sqrt{5})$), the Hilbert class field is generated over K by the real root α of the polynomial $X^3 - X - 1$. The singular modulus $j(\mathfrak{z}_D)$, which also generates this field, is equal to $-5^3\alpha^{12}(2\alpha-1)^3(3\alpha+2)^3$ and is a root of the much more complicated irreducible polynomial $H_{-23}(X) = X^3 + 3491750X^2 - 5151296875X + 12771880859375$, but using more detailed results from the theory of complex multiplication one can show that the number $2^{-1/2}e^{\pi i/24}\eta(\mathfrak{z}_D)/\eta(2\mathfrak{z}_D)$ also generates H, and this number turns out to be α itself! The improvement is even more dramatic for larger values of $|D|$ and is important in situations where one actually wants to compute a class field explicitly, as in the applications to factorization and primality testing mentioned in §6.4 below. ♡

♠ Solutions of Diophantine Equations

In §4.4 we discussed that one can parametrize an elliptic curve E over $\mathbb{Q}$ by modular functions $X(z)$, $Y(z)$, i.e., functions which are invariant under $\Gamma_0(N)$ for some N (the conductor of E) and identically satisfy the Weierstrass equation $Y^2 = X^3 + AX + B$ defining E. If K is an imaginary quadratic field with discriminant D prime to N and congruent to a square modulo $4N$ (this is equivalent to requiring that $\mathcal{O}_K = \mathcal{O}_D$ contains an ideal $\mathfrak{n}$ with $\mathcal{O}_D/\mathfrak{n} \cong \mathbb{Z}/N\mathbb{Z}$), then there is a canonical way to lift the $h(D)$ points of $\Gamma_1\backslash\mathfrak{H}$ of discriminant D to $h(D)$ points in the covering $\Gamma_0(N)\backslash\mathfrak{H}$, the so-called *Heegner points*. (The way is quite simple: if $D \equiv r^2 \pmod{4N}$, then one looks at points $\mathfrak{z}_Q$ with $Q(x,y) = Ax^2 + Bxy + Cy^2 \in \mathfrak{Q}_D$ satisfying $A \equiv 0 \pmod{N}$ and $B \equiv r \pmod{2N}$; there is exactly one $\Gamma_0(N)$-equivalence class of such points in each Γ_1-equivalence class of points of $\mathfrak{Z}_D$.) One can then show that the values of $X(z)$ and $Y(z)$ at the Heegner points behave exactly like the values of $j(z)$ at the points of $\mathfrak{Z}_D$, viz., these values all lie in the Hilbert class field H of K and are permuted simply transitively by the Galois group of H over K. It follows that we get $h(D)$ points $P_i = (X(\mathfrak{z}_i), Y(\mathfrak{z}_i))$ in E with coordinates in H which are permuted by $\mathrm{Gal}(H/K)$, and therefore that the sum $P_K = P_1 + \cdots + P_{h(D)}$ has coordinates in K (and in many cases even in $\mathbb{Q}$). This method of constructing potentially non-trivial rational solutions of Diophantine equations of genus 1 (the question of their actual non-triviality will be discussed briefly in §6.2) was invented by Heegner as part of his proof of the fact, mentioned above, that -163 is the smallest quadratic discriminant with class number 1 (his proof, which was rather sketchy at many points, was not accepted at the time by the mathematical community, but was later shown by Stark to be correct in all essentials) and has been used to construct non-trivial solutions of several classical Diophantine equations, e.g., to show that every prime number congruent to 5 or 7 (mod 8) is a "congruent number" (= the area of a right triangle with rational sides) and that every prime number congruent to 4 or 7 (mod 9) is a sum of two rational cubes (Sylvester's problem). ♡

6.2 Norms and Traces of Singular Moduli

A striking property of singular moduli is that they always are highly factorizable numbers. For instance, the value of $j(\mathfrak{z}_{-163})$ used in the first application in §6.1 is $-640320^3 = -2^{18}3^3 5^3 23^3 29^3$, and the value of $j(\mathfrak{z}_{-11}) = -32768$ is the 15th power of -2. As part of our investigation about the heights of Heegner points (see below), B. Gross and I were led to a formula which explains and generalizes this phenomenon. It turns out, in fact, that the right numbers to look at are not the values of the singular moduli themselves, but their *differences.* This is actually quite natural, since the definition of the modular function $j(z)$ involves an arbitrary choice of additive constant: any function $j(z) + C$ with $C \in \mathbb{Z}$ would have the same analytic and arithmetic properties as $j(z)$. The factorization of $j(\mathfrak{z})$, however, would obviously change completely if we replaced $j(z)$ by, say, $j(z) + 1$ or $j(z) - 744 = q^{-1} + \mathrm{O}(q)$ (the so-called "normalized Hauptmodul" of Γ_1), but this replacement would have no effect on the difference $j(\mathfrak{z}_1) - j(\mathfrak{z}_2)$ of two singular moduli. That the original singular moduli $j(\mathfrak{z})$ do nevertheless have nice factorizations is then due to the accidental fact that 0 itself is a singular modulus, namely $j(\mathfrak{z}_{-3})$, so that we can write them as differences $j(\mathfrak{z}) - j(\mathfrak{z}_{-3})$. (And the fact that the values of $j(\mathfrak{z})$ tend to be perfect cubes is then related to the fact that $\mathfrak{z}_{-3}$ is a fixed-point of order 3 of the action of Γ_1 on $\mathfrak{H}$.) Secondly, since the singular moduli are in general algebraic rather than rational integers, we should not speak only of their differences, but of the norms of their differences, and these norms will then also be highly factored. (For instance, the norms of the singular moduli $j(\mathfrak{z}_{-15})$ and $j(\mathfrak{z}_{-23})$, which are algebraic integers of degree 2 and 3 whose values were given in §6.1, are $-3^6 5^3 11^3$ and $-5^9 11^3 17^3$, respectively.)

If we restrict ourselves for simplicity to the case that the discriminants D_1 and D_2 of $\mathfrak{z}_1$ and $\mathfrak{z}_2$ are coprime, then the norm of $j(\mathfrak{z}_1) - j(\mathfrak{z}_2)$ depends only on D_1 and D_2 and is given by

$$J(D_1, D_2) \;=\; \prod_{\mathfrak{z}_1 \in \Gamma_1 \backslash \mathfrak{Z}_{D_1}} \; \prod_{\mathfrak{z}_2 \in \Gamma_1 \backslash \mathfrak{Z}_{D_2}} \bigl(j(\mathfrak{z}_1) \,-\, j(\mathfrak{z}_2)\bigr)\,. \tag{89}$$

(If $h(D_1) = 1$ then this formula reduces simply to $H_{D_2}(\mathfrak{z}_{D_1})$, while in general $J(D_1, D_2)$ is equal, up to sign, to the resultant of the two irreducible polynomials $H_{D_1}(X)$ and $H_{D_2}(X)$.) These are therefore the numbers which we want to study. We then have:

Theorem. *Let D_1 and D_2 be coprime negative discriminants. Then all prime factors of $J(D_1, D_2)$ are $\le \frac{1}{4}D_1D_2$. More precisely, any prime factors of $J(D_1, D_2)$ must divide $\frac{1}{4}(D_1D_2 - x^2)$ for some $x \in \mathbb{Z}$ with $|x| < \sqrt{D_1D_2}$ and $x^2 \equiv D_1D_2 \pmod 4$.*

There are in fact two proofs of this theorem, one analytic and one arithmetic. We give some brief indications of what their ingredients are, without defining all the terms occurring. In the analytic proof, one looks at the Hilbert modular group $\mathrm{SL}(2, \mathcal{O}_F)$ (see the notes by J. Bruinier in this volume) associated to the real quadratic field $F = \mathbb{Q}(\sqrt{D_1D_2})$ and constructs a certain Eisenstein series for this group, of weight 1 and with respect to the "genus character" associated to the decomposition of the discriminant of F as $D_1 \cdot D_2$. Then one restricts this form to the diagonal $z_1 = z_2$ (the story here is actually more complicated: the Eisenstein series in question is non-holomorphic and one has to take the holomorphic projection of its restriction) and makes use of the fact that there are no holomorphic modular forms of weight 2 on Γ_1. In the arithmetic proof, one uses that $p|J(D_1, D_2)$ if and only the CM elliptic curves with j-invariants $j(\mathfrak{z}_{D_1})$ and $j(\mathfrak{z}_{D_2})$ become isomorphic over $\overline{\mathbb{F}}_p$. Now it is known that the ring of endomorphisms of any elliptic curve over $\overline{\mathbb{F}}_p$ is isomorphic either to an order in a quadratic field or to an order in the (unique) quaternion algebra $B_{p,\infty}$ over $\mathbb{Q}$ ramified at p and at infinity. For the elliptic curve $\overline{E}$ which is the common reduction of the curves with complex multiplication by $\mathcal{O}_{D_1}$ and $\mathcal{O}_{D_2}$ the first alternative cannot occur, since a quadratic order cannot contain two quadratic orders coming from different quadratic fields, so there must be an order in $B_{p,\infty}$ which contains two elements α_1 and α_2 with square D_1 and D_2, respectively. Then the element $\alpha = \alpha_1\alpha_2$ also belongs to this order, and if x is its trace then $x \in \mathbb{Z}$ (because α is in an order and hence integral), $x^2 < N(\alpha) = D_1D_2$ (because $B_{p,\infty}$ is ramified at infinity), and $x^2 \equiv D_1D_2 \pmod p$ (because $B_{p,\infty}$ is ramified at p). This proves the theorem, except that we have lost a factor "4" because the elements α_i actually belong to the smaller orders $\mathcal{O}_{4D_i}$ and we should have worked with the elements $\alpha_i/2$ or $(1 + \alpha_i)/2$ (depending on the parity of D_i) in $\mathcal{O}_{D_i}$ instead.

The theorem stated above is actually only the qualitative version of the full result, which gives a complete formula for the prime factorization of $J(D_1, D_2)$. Assume for simplicity that D_1 and D_2 are fundamental. For each positive integer n of the form $\frac{1}{4}(D_1D_2 - x^2)$, we define a function ε from the set of divisors of n to $\{\pm 1\}$ by the requirement that ε is completely multiplicative (i.e., $\varepsilon\bigl(p_1^{r_1} \cdots p_s^{r_s}\bigr) = \varepsilon\bigl(p_1\bigr)^{r_1} \cdots \varepsilon\bigl(p_s\bigr)^{r_s}$ for any divisor $p_1^{r_1} \cdots p_s^{r_s}$ of n) and is given on primes $p|n$ by $\varepsilon(p) = \chi_{D_1}(p)$ if $p \nmid D_1$ and by $\varepsilon(p) = \chi_{D_2}(p)$ if $p \nmid D_2$, where χ_D is the Dirichlet character modulo D introduced at the beginning of §3.2. Notice that this makes sense: at least one of the two alternatives must hold, since $(D_1, D_2) = 1$ and p is prime, and if they both hold then the two definitions agree because D_1D_2 is then congruent to a non-zero square modulo p if p is odd and to an odd square modulo 8 if $p = 2$, so $\chi_{D_1}(p)\chi_{D_2}(p) = 1$. We then define $F(n)$ (still for n of the form

$\frac{1}{4}(D_1D_2 - x^2))$ by

$$F(n) \;=\; \prod_{d|n} d^{\varepsilon(n/d)}\,.$$

This number, which is a priori only rational since $\varepsilon(n/d)$ can be positive or negative, is actually integral and in fact *is always a power of a single prime number*: one can show easily that $\varepsilon(n) = -1$, so n contains an odd number of primes p with $\varepsilon(p) = -1$ and $2 \nmid \mathrm{ord}_p(n)$, and if we write

$$n \;=\; p_1^{2\alpha_1+1} \cdots p_r^{2\alpha_r+1} p_{r+1}^{2\beta_1} \cdots p_{r+s}^{2\beta_s} q_1^{\gamma_1} \cdots q_t^{\gamma_t}$$

with r odd and $\varepsilon(p_i) = -1$, $\varepsilon(q_j) = +1$, α_i, β_i, $\gamma_j \geq 0$, then

$$F(n) \;=\; \begin{cases} p^{(\alpha_1+1)(\gamma_1+1)\cdots(\gamma_t+1)} & \text{if } r \;=\; 1,\; p_1 \;=\; p\,, \\ \qquad 1 & \text{if } r \geq 3\,. \end{cases} \tag{90}$$

The complete formula for $J(D_1, D_2)$ is then

$$J(D_1, D_2)^{8/w(D_1)w(D_2)} \;=\; \prod_{\substack{x^2 < D_1D_2 \\ x^2 \equiv D_1D_2 \pmod 4}} F\biggl(\frac{D_1D_2 - x^2}{4}\biggr)\,. \tag{91}$$

As an example, for $D_1 = -7$, $D_2 = -43$ (both with class number one) this formula gives

$$\begin{aligned} J(-7,-43) \;&=\; \prod_{\substack{1 \leq x \leq 17 \\ x \text{ odd}}} F\biggl(\frac{301 - x^2}{4}\biggr) \\ &=\; F(75)F(73)F(69)F(63)F(55)F(45)F(33)F(19)F(3) \\ &=\; 3 \cdot 73 \cdot 3^2 \cdot 7 \cdot 5^2 \cdot 5 \cdot 3^2 \cdot 19 \cdot 3\,, \end{aligned}$$

and indeed $j(\mathfrak{z}_{-7}) - j(\mathfrak{z}_{-43}) = -3375 + 884736000 = 3^6 \cdot 5^3 \cdot 7 \cdot 19 \cdot 73$. This is the only instance I know of in mathematics where the prime factorization of a number (other than numbers like $n!$ which are defined as products) can be described in closed form.

♠ Heights of Heegner Points

This “application” is actually not an application of the result just described, but of the methods used to prove it. As already mentioned, the above theorem about differences of singular moduli was found in connection with the study of the height of Heegner points on elliptic curves. In the last “application” in §6.1 we explained what Heegner points are, first as points on the modular curve $X_0(N) = \Gamma_0(N)\backslash\mathfrak{H} \cup \{\text{cusps}\}$ and then via the modular parametrization as points on an elliptic curve E of conductor N. In fact we do not have to pass to an elliptic curve; the $h(D)$ Heegner points on $X_0(N)$ corresponding to complex multiplication by the order $\mathcal{O}_D$ are defined over the Hilbert class field H of $K = \mathbb{Q}(\sqrt{D})$ and we can add these points on the Jacobian $J_0(N)$ of $X_0(N)$ (rather than adding their images in E as before). The fact that they are permuted by $\mathrm{Gal}(H/K)$ means that this sum is a point $P_K \in J_0(N)$ defined over K (and sometimes even over $\mathbb{Q}$). The main question is whether this is a torsion point or not; if it is not, then we have an interesting solution of a Diophantine equation (e.g., a non-trivial point on an elliptic curve over $\mathbb{Q}$). In general, whether a point P on an elliptic curve or on an abelian variety (like $J_0(N)$) is torsion or not is measured by an invariant called the (global) *height* of P, which is always ≥ 0 and which vanishes if and only if P has finite order. This height is defined as the sum of *local heights*, some of which are "archimedean" (i.e., associated to the complex geometry of the variety and the point) and can be calculated as transcendental expressions involving Green's functions, and some of which are "non-archimedean" (i.e., associated to the geometry of the variety and the point over the p-adic numbers for some prime number p) and can be calculated arithmetically as the product of $\log p$ with an integer which measures certain geometric intersection numbers in characteristic p. Actually, the height of a point is the value of a certain positive definite quadratic form on the Mordell-Weil group of the variety, and we can also consider the associated bilinear form (the "height pairing"), which must then be evaluated for a pair of Heegner points. If $N = 1$, then the Jacobian $J_0(N)$ is trivial and therefore all global heights are automatically zero. It turns out that the archimedean heights are essentially the logarithms of the absolute values of the individual terms in the right-hand side of (89), while the p-adic heights are the logarithms of the various factors $F(n)$ on the right-hand side of (91) which are given by formula (90) as powers of the given prime number p. The fact that the global height vanishes is therefore equivalent to formula (91) in this case. If $N > 1$, then in general (namely, whenever $X_0(N)$ has positive genus) the Jacobian is non-trivial and the heights do not have to vanish identically. The famous conjecture of Birch and Swinnerton-Dyer, one of the seven million-dollar Clay Millennium Problems, says that the heights of points on an abelian variety are related to the values or derivatives of a certain L-function; more concretely, in the case of an elliptic curve $E/\mathbb{Q}$, the conjecture predicts that the order to which the L-series $L(E,s)$ vanishes at $s = 1$ is equal to the rank of the Mordell-Weil group $E(\mathbb{Q})$ and that the value of the first non-zero derivative of $L(E,s)$ at $s = 1$ is equal to a certain explicit expression involving the height pairings of a system of generators of $E(\mathbb{Q})$ with one another. The same kind of calculations as in the case $N = 1$ permitted a verification of this prediction in the case of Heegner points, the relevant derivative of the L-series being the first one:

Theorem. *Let E be an elliptic curve defined over $\mathbb{Q}$ whose L-series vanishes at $s = 1$. Then the height of any Heegner point is an explicit (and in general non-zero) multiple of the derivative $L'(E/\mathbb{Q}, 1)$.*

The phrase "in general non-zero" in this theorem means that for any given elliptic curve E with $L(E,1) = 0$ but $L'(E,1) \neq 0$ there are Heegner points whose height is non-zero and which therefore have infinite order in the Mordell-Weil group $E(\mathbb{Q})$. We thus get the following (very) partial statement in the direction of the full BSD conjecture:

Corollary. *If $E/\mathbb{Q}$ is an elliptic curve over $\mathbb{Q}$ whose L-series has a simple zero at $s = 1$, then the rank of $E(\mathbb{Q})$ is at least one.*

(Thanks to subsequent work of Kolyvagin using his method of "Euler systems" we in fact know that the rank is exactly equal to one in this case.) In the opposite direction, there are elliptic curves $E/\mathbb{Q}$ and Heegner points P in $E(\mathbb{Q})$ for which we know that the multiple occurring in the theorem is non-zero but where P can be checked directly to be a torsion point. In that case the theorem says that $L'(E/\mathbb{Q},1)$ must vanish and hence, if the L-series is known to have a functional equation with a minus sign (the sign of the functional equation can be checked algorithmically), that $L(E/\mathbb{Q},s)$ has a zero of order at least 3 at $s = 1$. This is important because it is exactly the hypothesis needed to apply an earlier theorem of Goldfeld which, assuming that such a curve is known, proves that the class numbers of $h(D)$ go to infinity in an effective way as $D \to -\infty$. Thus modular methods and the theory of Heegner points suffice to solve the nearly 200-year problem, due to Gauss, of showing that the set of discriminants $D < 0$ with a given class number is finite and can be determined explicitly, just as in Heegner's hands they had already sufficed to solve the special case when the given value of the class number was one. ♡

So far in this subsection we have discussed the *norms* of singular moduli (or more generally, the norms of their differences), but algebraic numbers also have *traces*, and we can consider these too. Now the normalization of j does matter; it turns out to be best to choose the normalized Hauptmodul $j_0(z) = j(z) - 744$. For every discriminant $D < 0$ we therefore define $T(D) \in \mathbb{Z}$ to be the trace of $j_0(\mathfrak{z}_D)$. This is the sum of the $h(D)$ singular moduli $j_0(\mathfrak{z}_{D,i})$, but just as in the case of the Hurwitz class numbers it is better to use the modified trace $T^*(d)$ which is defined as the sum of the values of $j_0(z)$ at all points $z \in \Gamma_1\backslash\mathfrak{H}$ satisfying a quadratic equation, primitive or not, of discriminant D (and with the points i and ρ, if they occur at all, being counted with multiplicity $1/2$ or $1/3$ as usual), i.e., $T^*(D) = \sum_{r^2|D} T(D/r^2)/(\frac{1}{2}w(D/r^2))$ with the same conventions as in the definition of $h^*(D)$. The result, quite different from (and much easier to prove than) the formula for the norms, is that these numbers $T^*(D)$ are the coefficients of a modular form. Specifically, if we denote by $g(z)$ the meromorphic modular form $\theta_M(z)E_4(4z)/\eta(4z)^6$ of weight $3/2$, where $\theta_M(z)$ is the Jacobi theta function $\sum_{n\in\mathbb{Z}}(-1)^n q^{n^2}$ as in §3.1, then we have:

Theorem. *The Fourier expansion of $g(z)$ is given by*

$$g(z) \;=\; q^{-1} \,-\, 2 \,-\, \sum_{d>0} T^*(-d)\, q^d\,. \tag{92}$$

We can check the first few coefficients of this by hand: the Fourier expansion of g begins $q^{-1} - 2 + 248\,q^3 - 492\,q^4 + 4119\,q^7 - \cdots$ and indeed we have $T^*(-3) = \frac{1}{3}(0 - 744) = -248$, $T^*(-4) = \frac{1}{2}(1728 - 744) = 492$, and $T^*(-7) = -3375 - 744 = -4119$.

The proof of this theorem, though somewhat too long to be given here, is fairly elementary and is essentially a refinement of the method used to prove the Hurwitz-Kronecker class number relations (87) and (88). More precisely, (87) was proved by comparing the degrees on both sides of (86), and these in turn were computed by looking at the most negative exponent occurring in the q-expansion of the two sides when X is replaced by $j(z)$; in particular, the number $\sigma_1^+(m)$ came from the calculation in Proposition 24 of the most negative power of q in $\Psi_m(j(z), j(z))$. If we look instead at the *next* coefficient (i.e., that of $q^{-\sigma_1^+(m)+1}$), then we find $-\sum_{t^2<4m} T^*(t^2 - 4m)$ for the right-hand side of (86), because the modular functions $H_D(j(z))$ and $H_D^*(j(z))$ have Fourier expansions beginning $q^{-h(D)}(1 - T(D)q + \mathrm{O}(q^2))$ and $q^{-h^*(D)}(1 - T^*(D)q + \mathrm{O}(q^2))$, respectively, while by looking carefully at the "lower order terms" in the proof of Proposition 24 we find that the corresponding coefficient for the left-hand side of (86) vanishes for all m unless m or $4m+1$ is a square, when it equals $+4$ or -2 instead. The resulting identity can then be stated uniformly for all m as

$$\sum_{t\in\mathbb{Z}} T^*(t^2 - 4m) \;=\; 0 \qquad (m \ge 0)\,, \tag{93}$$

where we have artificially defined $T^*(0) = 2$, $T^*(1) = -1$, and $T^*(D) = 0$ for $D > 1$ (a definition made plausible by the result in the theorem we want to prove). By a somewhat more complicated argument involving modular forms of higher weight, we prove a second relation, analogous to (88):

$$\sum_{t\in\mathbb{Z}} (m - t^2)T^*(t^2 - 4m) \;=\; \begin{cases} \;\;1 & \text{if } m \;=\; 0\,, \\ 240\sigma_3(n) & \text{if } m \ge 1\,. \end{cases} \tag{94}$$

(Of course, the factor $m - t^2$ here could be replaced simply by $-t^2$, in view of (93), but (94) is more natural, both from the proof and by analogy with (88).) Now, just as in the discussion of the Hurwitz-Kronecker class number relations, the two formulas (93) and (94) suffice to determine all $T^*(D)$ by recursion. But in fact we can solve these equations directly, rather than recursively (though this was not done in the original paper). Write the right-hand side of (92) as $t_0(z) + t_1(z)$ where $t_0(z) = \sum_{m=0}^{\infty} T^*(-4m)\,q^{4m}$ and $t_1(z) = \sum_{m=0}^{\infty} T^*(1-4m)\,q^{4m-1}$, and define two unary theta series $\theta_0(z)$ and $\theta_1(z)$ ($= \theta(4z)$ and $\theta_F(4z)$ in the notation of §3.1) as the sums $\sum q^{t^2}$ with t ranging over the even or odd integers, respectively. Then (93) says that $t_0\theta_0 + t_1\theta_1 = 0$ and (94) says that $[t_0, \theta_0] + [t_1, \theta_1] = 4E_4(4z)$, where $[t_i, \theta_i] = \frac{3}{2}t_i\theta_i' - \frac{1}{2}t_i'\theta_i$ is the first Rankin–Cohen bracket of t_i (in weight 3/2) and θ_i. By taking a linear combination of the second equation with the derivative of the first, we deduce that $\begin{pmatrix}\theta_0(z) & \theta_1(z)\\ \theta_0'(z) & \theta_1'(z)\end{pmatrix}\begin{pmatrix}t_0(z)\\ t_1(z)\end{pmatrix} = 2\begin{pmatrix}0\\ E_4(4z)\end{pmatrix}$, and from this we can immediately solve for $t_0(z)$ and $t_1(z)$ and hence for their sum, proving the theorem.

♠ The Borcherds Product Formula

This is certainly not an "application" of the above theorem in any reasonable sense, since Borcherds's product formula is much deeper and more general and was proved earlier, but it turns out that there is a very close link and that one can even use this to give an elementary proof of Borcherds's formula in a special case. This special case is the beautiful product expansion

$$H_D^*(j(z)) \;=\; q^{-h^*(D)}\prod_{n=1}^{\infty}\bigl(1\,-\,q^n\bigr)^{A_D(n^2)}\,, \tag{95}$$

where $A_D(m)$ denotes the mth Fourier coefficient of a certain meromorphic modular form $f_D(z)$ (with Fourier expansion beginning $q^D + \mathrm{O}(q)$) of weight 1/2. The link is that one can prove in an elementary way a "duality formula" $A_D(m) = -B_m(|D|)$, where $B_m(n)$ is the coefficient of q^n in the Fourier expansion of a certain other meromorphic modular form $g_m(z)$ (with Fourier expansion beginning $q^{-m} + \mathrm{O}(1)$) of weight 3/2. For $m = 1$ the function g_m coincides with the g of the theorem above, and since (95) immediately implies that $T^*(D) = A_D(1)$ this and the duality imply (92). Conversely, by applying Hecke operators (in half-integral weight) in a suitable way, one can give a generalization of (92) to all functions g_{n^2}, and this together with the duality formula gives the complete formula (95), not just its subleading coefficient. ♡

6.3 Periods and Taylor Expansions of Modular Forms

In §6.1 we showed that the value of any modular function (with rational or algebraic Fourier coefficients; we will not always repeat this) at a CM point $\mathfrak{z}$ is algebraic. This is equivalent to saying that for any modular form $f(z)$, of weight k, the value of $f(\mathfrak{z})$ is an algebraic multiple of $\Omega_{\mathfrak{z}}^k$, where $\Omega_{\mathfrak{z}}$ depends on $\mathfrak{z}$ only, not on f or on k. Indeed, the second statement implies the first by specializing to $k = 0$, and the first implies the second by observing that if $f \in M_k$ and $g \in M_\ell$ then f^ℓ/g^k has weight 0 and is therefore algebraic at $\mathfrak{z}$, so that $f(\mathfrak{z})^{1/k}$ and $g(\mathfrak{z})^{1/\ell}$ are algebraically proportional. Furthermore, the number $\Omega_{\mathfrak{z}}$ is unchanged (at least up to an algebraic number, but it is only defined up to an algebraic number) if we replace $\mathfrak{z}$ by $M\mathfrak{z}$ for any $M \in M(2,\mathbb{Z})$ with positive determinant, because $f(Mz)/f(z)$ is a modular function, and since any two CM points which generate the same imaginary quadratic field are related in this way, this proves:

Proposition 26. *For each imaginary quadratic field K there is a number $\Omega_K \in \mathbb{C}^*$ such that $f(\mathfrak{z}) \in \overline{\mathbb{Q}} \cdot \Omega_K^k$ for all $\mathfrak{z} \in K \cap \mathfrak{H}$, all $k \in \mathbb{Z}$, and all modular forms f of weight k with algebraic Fourier coefficients.* □

To find Ω_K, we should compute $f(\mathfrak{z})$ for some special modular form f (of non-zero weight!) and some point or points $\mathfrak{z} \in K \cap \mathfrak{H}$. A natural choice for the modular form is $\Delta(z)$, since it never vanishes. Even better, to achieve weight 1, is its 12th root $\eta(z)^2$, and better yet is the function $\Phi(z) = \Im(z)|\,\eta(z)|^4$ (which at $\mathfrak{z} \in K \cap \mathfrak{H}$ is an algebraic multiple of Ω_K^2), since it is Γ_1-invariant. As for the choice of $\mathfrak{z}$, we can look at the CM points of discriminant D ($=$ discriminant of K), but since there are $h(D)$ of them and none should be preferred over the others (their j-invariants are conjugate algebraic numbers), the only reasonable choice is to multiply them all together and take the $h(D)$-th root – or rather the $h'(D)$-th root (where $h'(D)$ as previously denotes $h(D)/\frac{1}{2}w(D)$, i.e., $h'(D) = \frac{1}{3}$, $\frac{1}{2}$ or $h(K)$ for $D = -3$, $D = -4$, or $D < -4$), because the elliptic fixed points ρ and i of $\Gamma_1\backslash\mathfrak{H}$ are always to be counted with multiplicity $\frac{1}{3}$ and $\frac{1}{2}$, respectively. Surprisingly enough, the product of the invariants $\Phi(\mathfrak{z}_{D,i})$ ($i = 1, \ldots, h(D)$) can be evaluated in closed form:

Theorem. *Let K be an imaginary quadratic field of discriminant D. Then*

$$\prod_{\mathfrak{z} \in \Gamma_1 \backslash \mathfrak{Z}_D} \bigl(4\pi\sqrt{|D|}\,\Phi(\mathfrak{z})\bigr)^{2/w(D)} \;=\; \prod_{j=1}^{|D|-1} \Gamma\bigl(j/|D|\bigr)^{\chi_D(j)}\,, \tag{96}$$

where χ_D is the quadratic character associated to K and $\Gamma(x)$ is the Euler gamma function.

Corollary. *The number Ω_K in Proposition 26 can be chosen to be*

$$\Omega_K \;=\; \frac{1}{\sqrt{2\pi|D|}} \left(\prod_{j=1}^{|D|-1} \Gamma\left(\frac{j}{|D|}\right)^{\chi_D(j)} \right)^{1/2h'(D)} . \tag{97}$$

Formula (96), usually called the Chowla–Selberg formula, is contained in a paper published by S. Chowla and A. Selberg in 1949, but it was later noticed that it already appears in a paper of Lerch from 1897. We cannot give the complete proof here, but we describe the main idea, which is quite simple. The Dedekind zeta function $\zeta_K(s) = \sum N(\mathfrak{a})^{-s}$ (sum over all non-zero integral ideals of K) has two decompositions: an additive one as $\sum_{\mathcal{A}} \zeta_{K,\mathcal{A}}(s)$, where $\mathcal{A}$ runs over the ideal classes of K and $\zeta_{K,\mathcal{A}}(s)$ is the associated "partial zeta function" ($= \sum N(\mathfrak{a})^{-s}$ with $\mathfrak{a}$ running over the ideals in $\mathcal{A}$), and a multiplicative one as $\zeta(s)L(s,\chi_D)$, where $\zeta(s)$ denotes the Riemann zeta function and $L(s,\chi_D) = \sum_{n=1}^{\infty} \chi_D(n)n^{-s}$. Using these two decompositions, one can compute in two different ways the two leading terms of the Laurent expansion $\zeta_K(s) = \frac{A}{s-1} + B + \mathrm{O}(s-1)$ as $s \to 1$. The residue at $s = 1$ of $\zeta_{K,\mathcal{A}}(s)$ is independent of $\mathcal{A}$ and equals $\pi/\frac{1}{2}w(D)\sqrt{|D|}$, leading to Dirichlet's class number formula $L(1,\chi_D) = \pi\, h'(D)/\sqrt{|D|}$. (This is, of course, precisely the method Dirichlet used.) The constant term in the Laurent expansion of $\zeta_{K,\mathcal{A}}(s)$ at $s=1$ is given by the famous *Kronecker limit formula* and is (up to some normalizing constants) simply the value of $\log(\Phi(\mathfrak{z}))$, where $\mathfrak{z} \in K \cap \mathfrak{H}$ is the CM point corresponding to the ideal class $\mathcal{A}$. The Riemann zeta function has the expansion $(s-1)^{-1} - \gamma + \mathrm{O}(s-1)$ near $s=1$ ($\gamma =$ Euler's constant), and $L'(1,\chi_D)$ can be computed by a relatively elementary analytic argument and turns out to be a simple multiple of $\sum_j \chi_D(j)\, \log\Gamma(j/|D|)$. Combining everything, one obtains (96).

As our first "application," we mention two famous problems of transcendence theory which were solved by modular methods, one using the Chowla–Selberg formula and one using quasimodular forms.

♠ Two Transcendence Results

In 1976, G.V. Chudnovsky proved that for any $z \in \mathfrak{H}$, at least two of the three numbers $E_2(z)$, $E_4(z)$ and $E_6(z)$ are algebraically independent. (Equivalently, the field generated by all $f(z)$ with $f \in \widetilde{M}_*(\Gamma_1)^{\mathbb{Q}} = \mathbb{Q}[E_2,\, E_4,\, E_6]$ has transcendence degree at least 2.) Applying this to $z=i$, for which $E_2(z) = 3/\pi$, $E_4(z) = 3\,\Gamma(\frac14)^8/(2\pi)^6$ and $E_6(z) = 0$, one deduces immediately that $\Gamma(\frac14)$ is transcendental (and in fact algebraically independent of π). Twenty years later, Nesterenko, building on earlier work of Barré-Sirieix, Diaz, Gramain and Philibert, improved this result dramatically by showing that for any $z \in \mathfrak{H}$ at least three of the four numbers $e^{2\pi i z}$, $E_2(z)$, $E_4(z)$ and $E_6(z)$ are algebraically independent. His proof used crucially the basic properties of the ring $\widetilde{M}_*(\Gamma_1)^{\mathbb{Q}}$ discussed in §5, namely, that it is closed under differentiation and that each of its elements is a power series in $q = e^{2\pi i z}$ with rational coefficients of bounded denominator and polynomial growth. Specialized to $z=i$, Nesterenko's result implies that the three numbers π, e^{π} and $\Gamma(\frac14)$ are algebraically independent. The algebraic independence of π and e^{π} (even without $\Gamma(\frac14)$) had been a famous open problem. ♡

♠ Hurwitz Numbers

Euler's famous result of 1734 that $\zeta(2r)/\pi^{2r}$ is rational for every $r \ge 1$ can be restated in the form

$$\sum_{\substack{n\in\mathbb{Z}\\ n\neq 0}} \frac{1}{n^k} \;=\; (\text{rational number})\cdot \pi^k \qquad \text{for all } k \ge 2\,, \tag{98}$$

where the "rational number" of course vanishes for k odd since then the contributions of n and $-n$ cancel. This result can be obtained, for instance, by looking at the Laurent expansion of $\cot x$ near the origin. Hurwitz asked the corresponding question if one replaces $\mathbb{Z}$ in (98) by the ring $\mathbb{Z}[i]$ of Gaussian integers and, by using an elliptic function instead of a trigonometric one, was able to prove the corresponding assertion

$$\sum_{\substack{\lambda\in\mathbb{Z}[i]\\ \lambda\neq 0}} \frac{1}{\lambda^k} \;=\; \frac{H_k}{k!}\,\omega^k \qquad \text{for all } k\geq 3\,, \tag{99}$$

for certain rational numbers $H_4=\frac{1}{10}$, $H_8=\frac{3}{10}$, $H_{12}=\frac{567}{130}$, ... (here $H_k=0$ for $4\nmid k$ since then the contributions of λ and $-\lambda$ or of λ and $i\lambda$ cancel), where

$$\omega \;=\; 4\int_0^1 \frac{dx}{\sqrt{1-x^4}} \;=\; \frac{\Gamma(\frac14)^2}{\sqrt{2\pi}} \;=\; 5.24411\cdots\,.$$

Instead of using the theory of elliptic functions, we can see this result as a special case of Proposition 26, since the sum on the left of (99) is just the special value of the modular form $2G_k(z)$ defined in (10), which is related by (12) to the modular form $\mathbb{G}_k(z)$ with rational Fourier coefficients, at $z=i$ and ω is $2\pi\sqrt{2}$ times the Chowla–Selberg period $\Omega_{\mathbb{Q}(i)}$. Similar considerations, of course, apply to the sum $\sum\lambda^{-k}$ with λ running over the non-zero elements of the ring of integers (or of any other ideal) in any imaginary quadratic field, and more generally to the special values of L-series of Hecke "grossencharacters" which we will consider shortly. ♡

We now turn to the second topic of this subsection: the Taylor expansions (as opposed to simply the values) of modular forms at CM points. As we already explained in the paragraph preceding equation (58), the "right" Taylor expansion for a modular form f at a point $z\in\mathfrak{H}$ is the one occurring on the left-hand side of that equation, rather than the straight Taylor expansion of f. The beautiful fact is that, if z is a CM point, then after a renormalization by dividing by suitable powers of the period Ω_z, each coefficient of this expansion is an algebraic number (and in many cases even rational). This follows from the following proposition, which was apparently first observed by Ramanujan.

Proposition 27. *The value of $E_2^*(\mathfrak{z})$ at a CM point $\mathfrak{z}\in K\cap\mathfrak{H}$ is an algebraic multiple of Ω_K^2.*

Corollary. *The value of $\partial^n f(\mathfrak{z})$, for any modular form f with algebraic Fourier coefficients, any integer $n\geq 0$ and any CM point $\mathfrak{z}\in K\cap\mathfrak{H}$, is an algebraic multiple of Ω_K^{k+2n}, where k is the weight of f.*

Proof. We will give only a sketch, since the proof is similar to that already given for $j(z)$. For $M=\left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right)\in\mathcal{M}_m$ we define $(E_2^*|_2M)(z)=m(cz+d)^{-2}E_2^*(Mz)$ (= the usual slash operator for the matrix $m^{-1/2}M\in\mathrm{SL}(2,\mathbb{R})$). From formula (1) and the fact that $E_2^*(z)$ is a linear combination of $1/y$ and a holomorphic function, we deduce immediately that the difference $E_2^*-E_2^*|_2M$ is a holomorphic modular form of weight 2 (on the subgroup of finite index $M^{-1}\Gamma_1 M \cap \Gamma_1$ of Γ_1). It then follows by an argument similar to that in the proof of Proposition 23 that the function

$$P_m(z,X) \;:=\; \prod_{M\in\Gamma_1\backslash\mathcal{M}_m} \bigl(X \,+\, E_2^*(z) \,-\, (E_2^*|_2M)(z)\bigr) \qquad (z\in\mathfrak{H},\; X\in\mathbb{C})$$

is a polynomial in X (of degree $\sigma_1(m)$) whose coefficients are modular forms of appropriate weights on Γ_1 with rational coefficients. (For example, $P_2(z,X) \,=\, X^3 \,-\, \frac{3}{4}\,E_4(z)\,X \,+\, \frac{1}{4}\,E_6(z)$.) The algebraicity of $E_2^*(\mathfrak{z})/\Omega_K^2$ for a CM point $\mathfrak{z}$ now follows from Proposition 26, since if $M\notin\mathbb{Z}\cdot\mathrm{Id}_2$ fixes $\mathfrak{z}$ then $E_2^*(\mathfrak{z}) - (E_2^*|_2M)(\mathfrak{z})$ is a non-zero algebraic multiple of $E_2^*(\mathfrak{z})$. The corollary follows from (58) and the fact that each non-holomorphic derivative $\partial^n f(z)$ is a polynomial in $E_2^*(z)$ with coefficients that are holomorphic modular forms with algebraic Fourier coefficients, as one sees from equation (66).

Propositions 26 and 27 are illustrated for $\mathfrak{z}=\mathfrak{z}_D$ and $f = E_2^*$, E_4, E_6 and Δ in the following table, in which α in the penultimate row is the real root of $\alpha^3-\alpha-1=0$. Observe that the numbers in the final column of this table are all units; this is part of the general theory.

D	$\dfrac{\lvert D\rvert^{1/2}E_2^*(\mathfrak{z}_D)}{\Omega_D^2}$	$\dfrac{E_4(\mathfrak{z}_D)}{\Omega_D^4}$	$\dfrac{E_6(\mathfrak{z}_D)}{\lvert D\rvert^{1/2}\Omega_D^6}$	$\dfrac{\Delta(\mathfrak{z}_D)}{\Omega_D^{12}}$
-3	0	0	24	-1
-4	0	12	0	1
-7	3	15	27	-1
-8	4	20	28	1
-11	8	32	56	-1
-15	$6+3\sqrt{5}$	$15+12\sqrt{5}$	$42+\frac{63}{\sqrt{5}}$	$\frac{3-\sqrt{5}}{2}$
-19	24	96	216	-1
-20	$12+4\sqrt{5}$	$40+12\sqrt{5}$	$72+\frac{112}{\sqrt{5}}$	$\sqrt{5}-2$
-23	$\frac{7+11\alpha+12\alpha^2}{a^{1/3}}$	$5\alpha^{1/3}(6+4\alpha+\alpha^2)$	$\frac{469+1176\alpha+504\alpha^2}{23}$	$-\alpha^{-8}$
-24	$12+12\sqrt{2}$	$60+24\sqrt{2}$	$84+72\sqrt{2}$	$3-2\sqrt{2}$

In the paragraph preceding Proposition 17 we explained that the non-holomorphic derivatives $\partial^n f$ of a holomorphic modular form $f(z)$ are more natural and more fundamental than the holomorphic derivatives $D^n f$, because the Taylor series $\sum D^n f(z)\,t^n/n!$ represents f only in the disk $|z'-z| < \mathfrak{I}(z)$ whereas the series (58) represents f everywhere. The above corollary gives a second reason to prefer the $\partial^n f$: at a CM point $\mathfrak{z}$ they are monomials in the period $\Omega_{\mathfrak{z}}$ with algebraic coefficients, whereas the derivatives $D^n f(\mathfrak{z})$ are polynomials in $\Omega_{\mathfrak{z}}$ by equation (57) (in which there can be no cancellation since $\Omega_{\mathfrak{z}}$ is known to be transcendental). If we set

$$c_n \;=\; c_n(f,\mathfrak{z},\Omega) \;=\; \frac{\partial^n f(\mathfrak{z})}{\Omega^{k+2n}} \qquad (n \,=\, 0,\,1,\,2,\,\ldots)\,, \tag{100}$$

where Ω is a suitably chosen algebraic multiple of Ω_K, then the c_n (up to a possible common denominator which can be removed by multiplying f by a suitable integer) are even algebraic *integers*, and they still can be considered to be “Taylor coefficients of f at $\mathfrak{z}$” since by (58) the series $\sum_{n=0}^{\infty} c_n t^n/n!$ equals $(Ct+D)^{-k} f\big(\frac{At+B}{Ct+D}\big)$ where $\big(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\big) = \big(\begin{smallmatrix} -\bar{z} & z \\ -1 & 1 \end{smallmatrix}\big)\big(\begin{smallmatrix} 1/4\pi y\Omega & 0 \\ 0 & \Omega \end{smallmatrix}\big) \in \mathrm{GL}(2,\mathbb{C})$. These normalized Taylor coefficients have several interesting number-theoretical applications, as we will see below, and from many points of view are actually better number-theoretic invariants than the more familiar coefficients a_n of the Fourier expansion $f = \sum a_n q^n$. Surprisingly enough, they are also much easier to calculate: unlike the a_n, which are mysterious numbers (think of the Ramanujan function $\tau(n)$) and have to be calculated anew for each modular form, the Taylor coefficients $\partial^n f$ or c_n are always given by a simple recursive procedure. We illustrate the procedure by calculating the numbers $\partial^n E_4(i)$, but the method is completely general and works the same way for all modular forms (even of half-integral weight, as we will see in §6.4).

Proposition 28. *We have* $\partial^n E_4(i) = p_n(0) E_4(i)^{1+n/2}$, *with* $p_n(t) \in \mathbb{Z}[\frac{1}{6}][t]$ *defined recursively by*

$$p_0(t) = 1, \quad p_{n+1}(t) = \frac{t^2-1}{2} p_n'(t) - \frac{n+2}{6} t p_n(t) - \frac{n(n+3)}{144} p_{n-1}(t) \quad (n \ge 0)\,.$$

Proof. Since $E_2^*(i) = 0$, equation (66) implies that $\widetilde{f}_\partial(i,X) = \widetilde{f}_\vartheta(i,X)$ for any $f \in M_k(\Gamma_1)$, i.e., we have $\partial^n f(i) = \vartheta^{[n]} f(i)$ for all $n \ge 0$, where $\{\vartheta^{[n]} f\}_{n=0,1,2,\ldots}$ is defined by (64). We use Ramanujan's notations Q and R for the modular forms $E_4(z)$ and $E_6(z)$. By (54), the derivation ϑ sends Q and R to $-\frac{1}{3}R$ and $-\frac{1}{2}Q^2$, respectively, so ϑ acts on $M_*(\Gamma_1) = \mathbb{C}[Q,R]$ as $-\frac{R}{3}\frac{\partial}{\partial Q} - \frac{Q^2}{2}\frac{\partial}{\partial R}$. Hence $\vartheta^{[n]} Q = P_n(Q,R)$ for all n, where the polynomials $P_n(Q,R) \in \mathbb{Q}[Q,R]$ are given recursively by

$$P_0 \;=\; Q\,, \qquad P_{n+1} \;=\; -\,\frac{R}{3}\,\frac{\partial P_n}{\partial Q} \;-\; \frac{Q^2}{2}\,\frac{\partial P_n}{\partial R} \;-\; n(n+3)\,\frac{Q P_{n-1}}{144}\,.$$

Since $P_n(Q,R)$ is weighted homogeneous of weight $2n+4$, where Q and R have weight 4 and 6, we can write $P_n(Q,R)$ as $Q^{1+n/2} p_n(R/Q^{3/2})$ where p_n is a polynomial in one variable. The recursion for P_n then translates into the recursion for p_n given in the proposition.

The first few polynomials p_n are $p_1 = -\frac{1}{3}t$, $p_2 = \frac{5}{36}$, $p_3 = -\frac{5}{72}t$, $p_4 = \frac{5}{216}t^2 + \frac{5}{288}$, $\ldots$, giving $\partial^2 E_4(i) = \frac{5}{36} E_4(i)^2$, $\partial^4 E_4(i) = \frac{5}{288} E_4(i)^3$, etc. (The values for n odd vanish because i is a fixed point of the element $S \in \Gamma_1$ of order 2.) With the same method we find $\partial^n f(i) = q_n(0) E_4(i)^{(k+2n)/4}$ for any modular form $f \in M_k(\Gamma_1)$, with polynomials q_n satisfying the same recursion as p_n but with $n(n+3)$ replaced by $n(n+k-1)$ (and of course with a different initial value $q_0(t)$). If we consider a CM point $\mathfrak{z}$ other than i, then the method and result are similar but we have to use (68) instead of (66), where ϕ is a quasimodular form differing from E_2 by a (meromorphic) modular form of weight 2 and chosen so that $\phi^*(\mathfrak{z})$ vanishes. If we replace Γ_1 by some other group Γ, then the same method works in principle but we need an explicit description of the ring $M_*(\Gamma)$ to replace the description of $M_*(\Gamma_1)$ as $\mathbb{C}[Q, R]$ used above, and if the genus of the group is larger than 0 then the polynomials $p_n(t)$ have to be replaced by elements q_n of some fixed finite algebraic extension of $\mathbb{Q}(t)$, again satisfying a recursion of the form $q_{n+1} = Aq_n' + (nB+C)q_n + n(n+k-1)Dq_{n-1}$ with A, B, C and D independent of n. The general result says that the values of the non-holomorphic derivatives $\partial^n f(z_0)$ of any modular form at any point $z_0 \in \mathfrak{H}$ are given "quasi-recursively" as special values of a sequence of algebraic functions in one variable which satisfy a differential recursion.

♠ Generalized Hurwitz Numbers

In our last "application" we studied the numbers $G_k(i)$ whose values, up to a factor of 2, are given by the Hurwitz formula (99). We now discuss the meaning of the non-holomorphic derivatives $\partial^n G_k(i)$. From (55) we find

$$\partial_k\left(\frac{1}{(mz+n)^k}\right) = \frac{k}{2\pi i(z-\bar{z})} \cdot \frac{m\bar{z}+n}{(mz+n)^{k+1}}$$

and more generally

$$\partial_k^r\left(\frac{1}{(mz+n)^k}\right) = \frac{(k)_r}{(2\pi i(z-\bar{z}))^r} \cdot \frac{(m\bar{z}+n)^r}{(mz+n)^{k+r}}$$

for all $r \geq 0$, where $\partial_k^r = \partial_{k+2r-2} \circ \cdots \circ \partial_{k+2} \circ \partial_k$ and $(k)_r = k(k+1)\cdots(k+r-1)$ as in §5. Thus

$$\partial^n G_k(i) = \frac{(k)_n}{2\,(-4\pi)^n} \sum_{\substack{\lambda \in \mathbb{Z}[i] \\ \lambda \neq 0}} \frac{\bar{\lambda}^n}{\lambda^{k+n}} \qquad (n = 0, 1, 2, \ldots)$$

(and similarly for $\partial^n G_k(\mathfrak{z})$ for any CM point $\mathfrak{z}$, with $\mathbb{Z}[i]$ replaced by $\mathbb{Z}\mathfrak{z}+\mathbb{Z}$ and -4π by $-4\pi\Im(\mathfrak{z})$). If we observe that the class number of the field $K = \mathbb{Q}(i)$ is 1 and that any integral ideal of K can be written as a principal ideal (λ) for exactly four numbers $\lambda \in \mathbb{Z}[i]$, then we can write

$$\sum_{\substack{\lambda \in \mathbb{Z}[i] \\ \lambda \neq 0}} \frac{\bar{\lambda}^n}{\lambda^{k+n}} = \sum_{\substack{\lambda \in \mathbb{Z}[i] \\ \lambda \neq 0}} \frac{\bar{\lambda}^{k+2n}}{(\lambda\bar{\lambda})^{k+n}} = 4 \sum_{\mathfrak{a}} \frac{\psi_{k+2n}(\mathfrak{a})}{N(\mathfrak{a})^{k+n}}$$

where the sum runs over the integral ideals $\mathfrak{a}$ of $\mathbb{Z}[i]$ and $\psi_{k+2n}(\mathfrak{a})$ is defined as $\bar{\lambda}^{k+2n}$, where λ is any generator of $\mathfrak{a}$. (This is independent of λ if $k+2n$ is divisible by 4, and in the contrary case the sum vanishes.) The functions ψ_{k+2n} are called Hecke "grossencharacters" (the German original of this semi-anglicized word is "Größencharaktere", with five differences of spelling, and means literally "characters of size," referring to the fact that these characters, unlike the usual ideal class characters of finite order, depend on the size of the generator of a principal ideal) and their L-series $L_K(s,\psi_{k+2n}) = \sum_{\mathfrak{a}} \psi_{k+2n}(\mathfrak{a})N(\mathfrak{a})^{-s}$ are an important class of L-functions with known analytic continuation and functional equations. The above calculation shows that $\partial^n G_k(i)$ (or more generally the non-holomorphic derivatives of any holomorphic Eisenstein series at any CM point) are simple multiples of special values of these L-series at integral arguments, and Proposition 28 and its generalizations give us an algorithmic way to compute these values in closed form. ♡

6.4 CM Elliptic Curves and CM Modular Forms

In the introduction to §6, we defined elliptic curves with complex multiplication as quotients $E = \mathbb{C}/\Lambda$ where $\lambda\Lambda \subseteq \Lambda$ for some non-real complex number λ. In that case, as we have seen, the lattice Λ is homothetic to $\mathbb{Z}\mathfrak{z} + \mathbb{Z}$ for some CM point $\mathfrak{z} \in \mathfrak{H}$ and the singular modulus $j(E) = j(\mathfrak{z})$ is algebraic, so E has a model over $\overline{\mathbb{Q}}$. The map from E to E induced by multiplication by λ is also algebraic and defined over $\overline{\mathbb{Q}}$. For simplicity, we concentrate on those $\mathfrak{z}$ whose discriminant is one of the 13 values $-3, -4, -7, \ldots, -163$ with class number 1, so that $j(\mathfrak{z}) \in \mathbb{Q}$ and E (but not the complex multiplication) can be defined over $\mathbb{Q}$. For instance, the three elliptic curves

$$y^2 = x^3 + x\,, \qquad y^2 = x^3 + 1\,, \qquad y^2 = x^3 - 35x - 98 \tag{101}$$

have j-invariants 1728, 0 and -3375, corresponding to multiplication by the orders $\mathbb{Z}[i]$, $\mathbb{Z}\bigl[\frac{1+\sqrt{-3}}{2}\bigr]$, and $\mathbb{Z}\bigl[\frac{1+\sqrt{-7}}{2}\bigr]$, respectively. For the first curve (or more generally any curve of the form $y^2 = x^3 + Ax$ with $A \in \mathbb{Z}$) the multiplication by i corresponds to the obvious endomorphism $(x,y) \mapsto (-x, iy)$ of the curve, and similarly for the second curve (or any curve of the form $y^2 = x^3 + B$) we have the equally obvious endomorphism $(x,y) \mapsto (\omega x, y)$ of order 3, where ω is a non-trivial cube root of 1. For the third curve (or any of its "twists" $E : Cy^2 = x^3 - 35x - 98$) the existence of a non-trivial endomorphism is less obvious. One checks that the map

$$\phi\,:\,(x,y) \;\mapsto\; \biggl(\gamma^2\Bigl(x + \frac{\beta^2}{x+\alpha}\Bigr),\; \gamma^3\, y\,\Bigl(1 - \frac{\beta^2}{(x+\alpha)^2}\Bigr)\biggr)\,,$$

where $\alpha = (7+\sqrt{-7})/2$, $\beta = (7+3\sqrt{-7})/2$, and $\gamma = (1+\sqrt{-7})/4$, maps E to itself and satisfies $\phi(\phi(P)) - \phi(P) + 2P = 0$ for any point P on E, where the addition is with respect to the group law on the curve, so that we have a map from $\mathcal{O}_{-7}$ to the endomorphisms of E sending $\lambda = m\frac{1+\sqrt{-7}}{2} + n$ to the endomorphism $P \mapsto m\phi(P) + nP$.

The key point about elliptic curves with complex multiplication is that the number of their points over finite fields is given by a simple formula. For the three curves above this looks as follows. Recall that the number of points over $\mathbb{F}_p$ of an elliptic curve $E/\mathbb{Q}$ given by a Weierstrass equation

$y^2 = F(x) = x^3 + Ax + B$ equals $p+1-a_p$ where a_p (for p odd and not dividing the discriminant of F) is given by $-\sum_{x \,(\mathrm{mod}\, p)} \bigl(\frac{F(x)}{p}\bigr)$. For the three curves in (101) we have

$$\sum_{x\,(\mathrm{mod}\,p)} \left(\frac{x^3+x}{p}\right) \;=\; \begin{cases} 0 & \text{if } p \equiv 3 \pmod 4\,, \\ -2a & \text{if } p \;=\; a^2+4b^2\,,\; a \equiv 1 \pmod 4 \end{cases} \tag{102a}$$

$$\sum_{x\,(\mathrm{mod}\,p)} \left(\frac{x^3+1}{p}\right) \;=\; \begin{cases} 0 & \text{if } p \equiv 2 \pmod 3\,, \\ -2a & \text{if } p \;=\; a^2+3b^2\,,\; a \equiv 1 \pmod 3 \end{cases} \tag{102b}$$

$$\sum_{x\,(\mathrm{mod}\,p)} \left(\frac{x^3-35x-98}{p}\right) \;=\; \begin{cases} 0 & \text{if } (p/7) \;=\; -1\,, \\ -2a & \text{if } p \;=\; a^2+7b^2\,,\; (a/7) \;=\; 1\,. \end{cases} \tag{102c}$$

(For other D with $h(D) = 1$ we would get a formula for $a_p(E)$ as $\pm A$ where $4p = A^2 + |D|B^2$.) The proofs of these assertions, and of the more general statements needed when $h(D) > 1$, will be omitted since they would take us too far afield, but we give the proof of the first (due to Gauss), since it is elementary and quite pretty. We prove a slightly more general but less precise statement.

Proposition 29. *Let p be an odd prime and A an integer not divisible by p. Then*

$$\sum_{x\,(\mathrm{mod}\,p)} \left(\frac{x^3+Ax}{p}\right) \;=\; \begin{cases} 0 & \textit{if } p \equiv 3 \pmod 4\,, \\ \pm 2a & \textit{if } p \equiv 1 \pmod 4 \textit{ and } (A/p) \;=\; 1\,, \\ \pm 4b & \textit{if } p \equiv 1 \pmod 4 \textit{ and } (A/p) \;=\; -1\,, \end{cases} \tag{103}$$

where $|a|$ and $|b|$ in the second and third lines are defined by $p = a^2 + 4b^2$.

Proof. The first statement is trivial since if $p \equiv 3 \pmod 4$ then $(-1/p) = -1$ and the terms for x and $-x$ in the sum cancel, so we can suppose that $p \equiv 1 \pmod 4$. Denote the sum on the left-hand side of (103) by $s_p(A)$. Replacing x by rx with $r \not\equiv 0 \pmod p$ shows that $s_p(r^2 A) = \bigl(\frac{r}{p}\bigr)\, s_p(A)$, so the number $s_p(A)$ takes on only four values, say $\pm 2\alpha$ for $A = g^{4i}$ or $A = g^{4i+2}$ and $\pm 2\beta$ for $A = g^{4i+1}$ or $A = g^{4i+3}$, where g is a primitive root modulo p. (That $s_p(A)$ is always even is obvious by replacing x by $-x$.) Now we take the sum of the squares of $s_p(A)$ as A ranges over all integers modulo p, noting that $s_p(0) = 0$. This gives

$$\begin{aligned} 2(p-1)(\alpha^2+\beta^2) \;&=\; \sum_{A,\,x,\,y\,\in\,\mathbb{F}_p} \left(\frac{x^3+Ax}{p}\right)\left(\frac{y^3+Ay}{p}\right) \\ &=\; \sum_{x,\,y\,\in\,\mathbb{F}_p} \left(\frac{xy}{p}\right) \sum_{A\,\in\,\mathbb{F}_p} \left(\frac{(x^2+A)(y^2+A)}{p}\right) \\ &=\; \sum_{x,\,y\,\in\,\mathbb{F}_p} \left(\frac{xy}{p}\right) \bigl(-1 \,+\, p\,\delta_{x^2,y^2}\bigr) \;=\; 2p(p-1)\,, \end{aligned}$$

where in the last line we have used the easy fact that $\sum_{z\in\mathbb{F}_p}\bigl(\frac{z(z+r)}{p}\bigr)$ equals $p-1$ for $r \equiv 0 \pmod{p}$ and -1 otherwise. (Proof: the first statement is obvious; the substitution $z \mapsto rz$ shows that the sum is independent of r if $r \not\equiv 0$; and the sum of the values for all integers r mod p clearly vanishes.) Hence $\alpha^2 + \beta^2 = p$. It is also obvious that α is odd (for $(A/p) = +1$ there are $\frac{p-1}{2} - 1$ values of x between 0 and $p/2$ with $\bigl(\frac{x^3+Ax}{p}\bigr)$ non-zero and hence odd) and that β is even (for $(A/p) = -1$ all $\frac{p-1}{2}$ values of $\bigl(\frac{x^3+Ax}{p}\bigr)$ with $0 < x < p/2$ are odd), so $\alpha = \pm a$, $\beta = \pm 2b$ as asserted.

An almost exactly similar proof works for equation (102b): here one defines $s_p(B) = \sum_x \bigl(\frac{x^3+B}{p}\bigr)$ and observes that $s_p(r^3B) = \bigl(\frac{r}{p}\bigr)\, s_p(B)$, so that $s_p(B)$ takes on six values $\pm\alpha$, $\pm\beta$ and $\pm\gamma$ depending on the class of i (mod 6), where $p \equiv 1 \pmod{6}$ (otherwise all $s_p(B)$ vanish) and $B = g^i$ with g a primitive root modulo p. Summing $s_p(B)$ over all quadratic residues $B \pmod{p}$ shows that $\alpha+\beta+\gamma = 0$, and summing $s_p(B)^2$ over all B gives $\alpha^2+\beta^2+\gamma^2 = 6p$. These two equations imply that $\alpha \equiv \beta \equiv \gamma \not\equiv 0 \pmod{3}$ and that $4p = \alpha^2+3((\beta-\gamma)/3)^2$, which gives (102b) since it is easily seen that α is even and β and γ are odd. I do not know of any elementary proof of this sort for equation (102c) or for the similar identities corresponding to the other imaginary quadratic fields of class number 1. The reader is urged to try to find such a proof.

♠ Factorization, Primality Testing, and Cryptography

We mention briefly one "practical" application of complex multiplication theory. Many methods of modern cryptography depend on being able to identify very large prime numbers quickly or on being able to factor (or being relatively sure that no one else will be able to factor) very large composite numbers. Several methods involve the arithmetic of elliptic curves over finite fields, which yield finite groups in which certain operations are easily performed but not easily inverted. The difficulty of the calculations, and hence the security of the method, depends on the structure of the group of points of the curve over the finite fields, so one would like to be able to construct, say, examples of elliptic curves E over $\mathbb{Q}$ whose reduction modulo p for some very large prime p has an order which itself contains a very large prime factor. Since counting the points on $E(\mathbb{F}_p)$ directly is impractical when p is very large, it is essential here to know curves E for which the number a_p, and hence the cardinality of $E(\mathbb{F}_p)$, is known *a priori*, and the existence of closed formulas like the ones in (102) implies that the curves with complex multiplication are suitable. In practice one wants the complex multiplication to be by an order in a quadratic field which is not too big but also not too small. For this purpose one needs effective ways to construct the Hilbert class fields (which is where the needed singular moduli will lie) efficiently, and here again the methods mentioned in 6.1, and in particular the simplifications arising by replacing the modular function $j(z)$ by better modular functions, become relevant. For more information, see the bibliography. ♡

We now turn from elliptic curves with complex multiplication to an important and related topic, the so-called CM modular forms. Formula (102a) says that the coefficient a_p of the L-series $L(E,s) = \sum a_n n^{-s}$ of the elliptic curve $E: y^2 = x^3 + x$ for a prime p is given by $a_p = \mathrm{Tr}(\lambda) = \lambda + \bar{\lambda}$, where λ is one of the two numbers of norm p in $\mathbb{Z}[i]$ of the form $a + bi$ with $a \equiv 1 \pmod{4}$, $b \equiv 0 \pmod{2}$ (or is zero if there is no such λ). Now using the multiplicative properties of the a_n we find that the full L-series is given by

$$L(E,\,s) \;=\; L_K(s,\psi_1)\,, \tag{104}$$

where $L_K(s,\psi_1) = \sum_{0\neq\mathfrak{a}\subseteq\mathbb{Z}[i]} \psi_1(\mathfrak{a}) N(\mathfrak{a})^{-s}$ is the L-series attached as in §6.3 to the field $K = \mathbb{Q}(i)$ and the grossencharacter ψ_1 defined by

$$\psi_1(\mathfrak{a}) \;=\; \begin{cases} 0 & \text{if } 2 \mid N(\mathfrak{a})\,, \\ \lambda & \text{if } 2 \nmid N(\mathfrak{a})\,,\ \mathfrak{a} \;=\; (\lambda),\ \lambda \in 4\mathbb{Z} + 1 + 2\mathbb{Z}\,i\,. \end{cases} \tag{105}$$

In fact, the L-series of the elliptic curve E belongs to *three* important classes of L-series: it is an L-function coming from algebraic geometry (by definition), the L-series of a generalized character of an algebraic number field (by (104)), and also the L-series of a modular form, namely $L(E,s) = L(f_E,s)$ where

$$f_E(z) \;=\; \sum_{0\neq\mathfrak{a}\subseteq\mathbb{Z}[i]} \psi_1(\mathfrak{a})\,q^{N(\mathfrak{a})} \;=\; \sum_{\substack{a\equiv1 \pmod{4} \\ b\equiv0 \pmod{2}}} a\,q^{a^2+b^2} \tag{106}$$

which is a theta series of the type mentioned in the final paragraph of §3, associated to the binary quadratic form $Q(a,b) = a^2 + b^2$ and the spherical polynomial $P(a,b) = a$ (or $a+ib$) of degree 1. The same applies, with suitable modifications, for any elliptic curve with complex multiplication, so that the Taniyama-Weil conjecture, which says that the L-series of an elliptic curve over $\mathbb{Q}$ should coincide with the L-series of a modular form of weight 2, can be seen explicitly for this class of curves (and hence was known for them long before the general case was proved).

The modular form f_E defined in (106) can also be denoted θ_{ψ_1}, where ψ_1 is the grossencharacter (105). More generally, the L-series of the powers $\psi_d = \psi_1^d$ of ψ_1 are the L-series of modular forms, namely $L_K(\psi_d,s) = L(\theta_{\psi_d},s)$ where

$$\theta_{\psi_d}(z) \;=\; \sum_{0\neq\mathfrak{a}\subseteq\mathbb{Z}[i]} \psi_d(\mathfrak{a})\,q^{N(\mathfrak{a})} \;=\; \sum_{\substack{a\equiv1 \pmod{4} \\ b\equiv0 \pmod{2}}} (a+ib)^d\,q^{a^2+b^2}\,, \tag{107}$$

which is a modular form of weight $d+1$. Modular forms constructed in this way – i.e., theta series associated to a binary quadratic form $Q(a,b)$ and a spherical polynomial $P(a,b)$ of arbitrary degree d – are called *CM modular forms*, and have several remarkable properties. First of all, they are always linear combinations of the theta series $\theta_\psi(z) = \sum_{\mathfrak{a}} \psi(\alpha)\,q^{N(\mathfrak{a})}$ associated to some grossencharacter ψ of an imaginary quadratic field. (This is because any positive definite binary quadratic form over $\mathbb{Q}$ is equivalent to the norm form $\lambda \mapsto \lambda\bar{\lambda}$ on an ideal in an imaginary quadratic field, and since the Laplace operator associated to this form is $\frac{\partial^2}{\partial\lambda\,\partial\bar{\lambda}}$, the only spherical polynomials of degree d are linear combinations of λ^d and $\bar{\lambda}^d$.) Secondly, since the L-series $L(\theta_\psi, s) = L_K(\psi, s)$ has an Euler product, the modular forms θ_ψ attached to grossencharacters are always Hecke eigenforms, and this is the *only* infinite family of Hecke eigenforms, apart from Eisenstein series, which are known explicitly. (Other eigenforms do not appear to have any systematic rule of construction, and even the number fields in which their Fourier coefficients lie are totally mysterious, an example being the field $\mathbb{Q}(\sqrt{144169})$ of coefficients of the cusp form of level 1 and weight 24 mentioned in §4.1.) Thirdly, sometimes modular forms constructed by other methods turn out to be of CM type, leading to new identities. For instance, in four cases the modular form defined in (107) is a product of eta-functions: $\theta_{\psi_0}(z) = \eta(8z)^4/\eta(4z)^2$, $\theta_{\psi_1}(z) = \eta(8z)^8/\eta(4z)^2\eta(16z)^2$, $\theta_{\psi_2}(z) = \eta(4z)^6$, $\theta_{\psi_4}(z) = \eta(4z)^{14}/\eta(8z)^4$. More generally, we can ask which eta-products are of CM type. Here I do not know the answer, but for pure powers there is a complete result, due to Serre. The fact that both $\eta(z) = \sum\limits_{n\equiv 1 \pmod 6} (-1)^{(n-1)/6} q^{n^2/24}$ and $\eta(z)^3 = \sum\limits_{n\equiv 1 \pmod 4} n\, q^{n^2/8}$ are unary theta series implies that each of the functions η^2, η^4 and η^6 is a binary theta series and hence a modular form of CM type (the function $\eta(z)^6$ equals $\theta_{\psi_2}(z/4)$, as we just saw, and $\eta(z)^4$ equals $f_E(z/6)$, where E is the second curve in (101)), but there are other, less obvious, examples, and these can be completely classified:

Theorem (Serre). *The function $\eta(z)^n$ (n even) is a CM modular form for $n = 2$, 4, 6, 8, 10, 14 or 26 and for no other value of n.*

Finally, the CM forms have another property called "lacunarity" which is not shared by any other modular forms. If we look at (106) or (107), then we see that the only exponents which occur are sums of two squares. By the theorem of Fermat proved in §3.1, only half of all primes (namely, those congruent to 1 modulo 4) have this property, and by a famous theorem of Landau, only $\mathrm{O}(x/(\log x)^{1/2})$ of the integers $\le x$ do. The same applies to any other CM form and shows that 50% of the coefficients a_p (p prime) vanish if the form is a Hecke eigenform and that 100% of the coefficients a_n ($n \in \mathbb{N}$) vanish for any CM form, eigenform or not. Another difficult theorem, again due to Serre, gives the converse statements:

Theorem (Serre). *Let $f = \sum a_n q^n$ be a modular form of integral weight ≥ 2. Then:*

1. *If f is a Hecke eigenform, then the density of primes p for which $a_p \ne 0$ is equal to $1/2$ if f corresponds to a grossencharacter of an imaginary quadratic field, and to 1 otherwise.*
2. *The number of integers $n \le x$ for which $a_n \ne 0$ is $\mathrm{O}(x/\sqrt{\log x})$ as $x \to \infty$ if f is of CM type and is larger than a positive multiple of x otherwise.*

♠ Central Values of Hecke L-Series

We just saw above that the Hecke L-series associated to grossencharacters of degree d are at the same time the Hecke L-series of CM modular forms of weight $d+1$, and also, if $d=1$, sometimes the L-series of elliptic curves over $\mathbb{Q}$. On the other hand, the Birch–Swinnerton-Dyer conjecture predicts that the value of $L(E,1)$ for any elliptic curve E over $\mathbb{Q}$ (not just one with CM) is related as follows to the arithmetic of E: it vanishes if and only if E has a rational point of infinite order, and otherwise is (essentially) a certain period of E multiplied by the order of a mysterious group Ш, the Tate–Shafarevich group of E. Thanks to the work of Kolyvagin, one knows that this group is indeed finite when $L(E,1) \neq 0$, and it is then a standard fact (because Ш admits a skew-symmetric non-degenerate pairing with values in $\mathbb{Q}/\mathbb{Z}$) that the order of Ш is a perfect square. In summary, the value of $L(E,1)$, normalized by dividing by an appropriate period, is always a perfect square. This suggests looking at the central point (= point of symmetry with respect to the functional equation) of other types of L-series, and in particular of L-series attached to grossencharacters of higher weights, since these can be normalized in a nice way using the Chowla–Selberg period (97), to see whether these numbers are perhaps also always squares.

An experiment to test this idea was carried out over 25 years ago by B. Gross and myself and confirmed this expectation. Let K be the field $\mathbb{Q}(\sqrt{-7})$ and for each $d \geq 1$ let $\psi_d = \psi_1^d$ be the grossencharacter of K which sends an ideal $\mathfrak{a} \subseteq \mathcal{O}_K$ to λ^d if $\mathfrak{a}$ is prime to $\mathfrak{p}_7 = (\sqrt{-7})$ and to 0 otherwise, where λ in the former case is the generator of $\mathfrak{a}$ which is congruent to a square modulo $\mathfrak{p}_7$. For $d=1$, the L-series of ψ_d coincides with the L-series of the third elliptic curve in (101), while for general odd values of d it is the L-series of a modular form of weight $d+1$ and trivial character on $\Gamma_0(49)$. The L-series $L(\psi_{2m-1},s)$ has a functional equation sending s to $2m-s$, so the point of symmetry is $s=m$. The central value has the form

$$L(\psi_{2m-1},m) \;=\; \frac{2A_m}{(m-1)!}\,\Bigl(\frac{2\pi}{\sqrt{7}}\Bigr)^m\,\Omega_K^{2m-1} \tag{108}$$

where

$$\Omega_K \;=\; \sqrt[4]{7}\,\left|\eta\Bigl(\frac{1+\sqrt{-7}}{2}\Bigr)\right|^2 \;=\; \frac{\Gamma(\frac{1}{7})\Gamma(\frac{2}{7})\Gamma(\frac{4}{7})}{4\pi^2}$$

is the Chowla–Selberg period attached to K and (it turns out) $A_m \in \mathbb{Z}$ for all $m>1$. (We have $A_1 = \frac{1}{4}$.) The numbers A_m vanish for m even because the functional equation of $L(\psi_{2m-1},s)$ has a minus sign in that case, but the numerical computation suggested that the others were indeed all perfect squares: $A_3 = A_5 = 1^2$, $A_7 = 3^2$, $A_9 = 7^2$, …, $A_{33} = 44762286327255^2$. Many years later, in a paper with Fernando Rodriguez Villegas, we were able to confirm this prediction:

Theorem. *The integer A_m is a square for all $m>1$. More precisely, we have $A_{2n+1}=b_n(0)^2$ for all $n\ge 0$, where the polynomials $b_n(x)\in\mathbb{Q}[x]$ are defined recursively by $b_0(x)=\frac12$, $b_1(x)=1$ and $21\,b_{n+1}(x)=-(x-7)(64x-7)b_n'(x)+(32nx-56n+42)b_n(x)-2n(2n-1)(11x+7)b_{n-1}(x)$ for $n\ge 1$.*

The proof is too complicated to give here, but we can indicate the main idea. The first point is that the numbers A_m themselves can be computed by the method explained in §6.3. There we saw (for the full modular group, but the method works for higher level) that the value at $s=k+n$ of the L-series of a grossencharacter of degree $d=k+2n$ is essentially equal to the nth non-holomorphic derivative of an Eisenstein series of weight k at a CM point. Here we want $d=2m-1$ and $s=m$, so $k=1$, $n=m-1$. More precisely, one finds that $L(\psi_{2m-1},m)=\frac{(2\pi/\sqrt7)^m}{(m-1)!}\partial^{m-1}\mathbb{G}_{1,\varepsilon}(\mathfrak{z}_7)$, where $\mathbb{G}_{1,\varepsilon}$ is the Eisenstein series

$$\mathbb{G}_{1,\varepsilon} \;=\; \frac12+\sum_{n=1}^{\infty}\biggl(\sum_{d|n}\varepsilon(n)\biggr)q^n \;=\; \frac12+q+2q^2+3q^4+q^7+\cdots$$

of weight 1 associated to the character $\varepsilon(n)=\left(\frac{n}{7}\right)$ and $\mathfrak{z}_7$ is the CM point $\frac12\left(1+\frac{i}{\sqrt7}\right)$. These coefficients can be obtained by a quasi-recursion like the one in Proposition 28 (though it is more complicated here because the analogues of the polynomials $p_n(t)$ are now elements in a quadratic extension of $\mathbb{Q}[t]$), and therefore are very easy to compute, but this does not explain why they are squares. To see this, we first observe that the Eisenstein series $\mathbb{G}_{1,\varepsilon}$ is one-half of the binary theta series $\Theta(z)=\sum_{m,\,n\in\mathbb{Z}}q^{m^2+mn+2n^2}$. In his thesis, Villegas proved a beautiful formula expressing certain linear combinations of values of binary theta series at CM points as the squares of linear combinations of values of unary theta series at other CM points. The same turns out to be true for the higher non-holomorphic derivatives, and in the case at hand we find the remarkable formula

$$\partial^{2n}\Theta(\mathfrak{z}_7) \;=\; 2^{2n}7^{2n+1/4}\left|\partial^n\theta_2(\mathfrak{z}_7^*)\right|^2 \qquad \text{for all } n\ge 0\,, \tag{109}$$

where $\theta_2(z)=\sum_{n\in\mathbb{Z}}q^{(n+1/2)^2/2}$ is the Jacobi theta-series defined in (32) and $\mathfrak{z}_7^*=\frac12(1+i\sqrt7)$. Now the values of the non-holomorphic derivatives $\partial^n\theta_2(\mathfrak{z}_7^*)$ can be computed quasi-recursively by the method explained in 6.3. The result is the formula given in the theorem.

Remarks 1. By the same method, using that every Eisenstein series of weight 1 can be written as a linear combination of binary theta series, one can show that the correctly normalized central values of all L-series of grossencharacters of odd degree are perfect squares.

2. Identities like (109) seem very surprising. I do not know of any other case in mathematics where the Taylor coefficients of an analytic function at some point are in a non-trivial way the squares of the Taylor coefficients of another analytic function at another point.

3. Equation (109) is a special case of a yet more general identity, proved in the same paper, which expresses the non-holomorphic derivative $\partial^{2n}\theta_\psi(\mathfrak{z}_0)$ of the CM modular form associated to a grossencharacter ψ of degree $d = 2r$ as a simple multiple of $\partial^n\theta_2(\mathfrak{z}_1)\,\partial^{n+r}\theta_2(\mathfrak{z}_2)$, where $\mathfrak{z}_0$, $\mathfrak{z}_1$ and $\mathfrak{z}_2$ are CM points belonging to the quadratic field associated to ψ. ♡

We end with an application of these ideas to a classical Diophantine equation.

♠ Which Primes are Sums of Two Cubes?

In §3 we gave a modular proof of Fermat's theorem that a prime number can be written as a sum of two squares if and only if it is congruent to 1 modulo 4. The corresponding question for cubes was studied by Sylvester in the 19th century. For squares the answer is the same whether one considers integer or rational squares (though in the former case the representation, when it exists, is unique and in the latter case there are infinitely many), but for cubes one only considers the problem over the rational numbers because there seems to be no rule for deciding which numbers have a decomposition into integral cubes. The question therefore is equivalent to asking whether the Mordell-Weil group $E_p(\mathbb{Q})$ of the elliptic curve $E_p : x^3 + y^3 = p$ is non-trivial. Except for $p = 2$, which has the unique decomposition $1^3 + 1^3$, the group $E_p(\mathbb{Q})$ is torsion-free, so that if there is even one rational solution there are infinitely many. An equivalent question is therefore: for which primes p is the rank r_p of $E_p(\mathbb{Q})$ greater than 0?

Sylvester's problem was already mentioned at the end of §6.1 in connection with Heegner points, which can be used to construct non-trivial solutions if $p \equiv 4$ or $7 \mod 9$. Here we consider instead an approach based on the Birch–Swinnerton-Dyer conjecture, according to which $r_p > 0$ if and only if the L-series of E_p vanishes at $s = 1$. By a famous theorem of Coates and Wiles, one direction of this conjecture is known: if $L(E_p, 1) \neq 0$ then $E_p(\mathbb{Q})$ has rank 0. The question we want to study is therefore: when does $L(E_p, 1)$ vanish? For five of the six possible congruence classes for $p \pmod 9$ (we assume that $p > 3$, since $r_2 = r_3 = 0$) the answer is known. If $p \equiv 4$, 7 or 8 $\pmod 9$, then the functional equation of $L(E_p, s)$ has a minus sign, so $L(E_p, 1) = 0$ and r_p is expected (and, in the first two cases, known) to be ≥ 1; it is also known by an "infinite descent" argument to be ≤ 1 in these cases. If $p \equiv 2$ or 5 $\pmod 9$, then the functional equation has a plus sign and $L(E_p, 1)$ divided by a suitable period is $\equiv 1 \pmod 3$ and hence $\neq 0$, so by the Coates-Wiles theorem these primes can never be sums of two cubes, a result which can also be proved in an elementary way by descent. In the remaining case $p \equiv 1 \pmod 9$, however, the answer can vary: here the functional equation has a plus sign, so $\mathrm{ord}_{s=1}L(E_p, s)$ is even and r_p is also expected to be even, and descent gives $r_p \leq 2$, but both cases $r_p = 0$ and $r_p = 2$ occur. The following result, again proved jointly with F. Villegas, gives a criterion for these primes.

Theorem. *Define a sequence of numbers* $c_0 = 1$, $c_1 = 2$, $c_2 = -152$, $c_3 = 6848, \ldots$ *by* $c_n = s_n(0)$ *where* $s_0(x) = 1$, $s_1(x) = 3x^2$ *and* $s_{n+1}(x) = (1-8x^3)s_n'(x)+(16n+3)x^2 s_n(x)-4n(2n-1)x s_{n-1}(x)$ *for* $n \geq 1$. *If* $p = 9k+1$ *is prime, then* $L(E_p, 1) = 0$ *if and only if* $p|c_k$.

For instance, the numbers $c_2 = -152$ and $c_4 = -8103296$ are divisible by $p = 19$ and $p = 37$, respectively, so $L(E_p, 1)$ vanishes for these two primes (and indeed $19 = 3^3 + (-2)^3$, $37 = 4^3 + (-3)^3$), whereas $c_8 = 532650564250569441280$ is not divisible by 73, which is therefore not a sum of two rational cubes. Note that the numbers c_n grow very quickly (roughly like n^{3n}), but to apply the criterion for a given prime number $p = 9k + 1$ one need only compute the polynomials $s_n(x)$ modulo p for $0 \leq n \leq k$, so that no large numbers are required.

We again do not give the proof of this theorem, but only indicate the main ingredients. The central value of $L(E_p, s)$ for $p \equiv 1 \pmod 9$ is given by $L(E_p, 1) = \dfrac{3\,\Omega}{\sqrt[3]{p}}\, S_p$ where $\Omega = \dfrac{\Gamma(\frac{1}{3})^3}{2\pi\sqrt{3}}$ and S_p is an integer which is supposed to be a perfect square (namely the order of the Tate-Shafarevich group of E_p if $r_p = 0$ and 0 if $r_p > 0$). Using the methods from Villegas's thesis, one can show that this is true and that both S_p and its square root can be expressed as the traces of certain algebraic numbers defined as special values of modular functions at CM points: $S_p = \mathrm{Tr}(\alpha_p) = \mathrm{Tr}(\beta_p)^2$, where $\alpha_p = \dfrac{\sqrt[3]{p}}{54}\,\dfrac{\Theta(p\mathfrak{z}_0)}{\Theta(\mathfrak{z}_0)}$ and $\beta_p = \dfrac{\sqrt[6]{p}}{\sqrt{\pm 12}}\,\dfrac{\eta(p\mathfrak{z}_1)}{\eta(\mathfrak{z}_1/p)}$ with $\Theta(z) = \sum\limits_{m,n} q^{m^2+mn+n^2}$, $\mathfrak{z}_0 = \dfrac{1}{2} + \dfrac{i}{6\sqrt{3}}$ and $\mathfrak{z}_1 = \dfrac{r+\sqrt{-3}}{2}$ ($r \in \mathbb{Z}$, $r^2 \equiv -3 \pmod{4p}$). This gives an explicit formula for $L(E_p, 1)$, but it is not very easy to compute since the numbers α_p and β_p have large degree ($18k$ and $6k$, respectively if $p = 9k + 1$) and lie in a different number field for each prime p. To obtain a formula in which everything takes place over $\mathbb{Q}$, one observes that the L-series of E_p is the L-series of a cubic twist of a grossencharacter ψ_1 of $K = \mathbb{Q}(\sqrt{-3})$ which is independent of p. More precisely, $L(E_p, s) = L(\chi_p\psi_1, s)$ where ψ_1 is defined (just like the grossencharacters for $\mathbb{Q}(i)$ and $\mathbb{Q}(\sqrt{-7})$ defined in (105) and in the previous "application") by $\psi_1(\mathfrak{a}) = \lambda$ if $\mathfrak{a} = (\lambda)$ with $\lambda \equiv 1 \pmod 3$ and $\psi_1(\alpha) = 0$ if $3|N(\mathfrak{a})$, and χ_p is the cubic character which sends $\mathfrak{a} = (\lambda)$ to the unique cube root of unity in K which is congruent to $(\bar{\lambda}/\lambda)^{(p-1)/3}$ modulo p. This means that formally the L-value $L(E_p, 1)$ for $p = 9k + 1$ is congruent modulo p to the central value $L(\psi^{12k+1}, 6k+1)$. Of course both of these numbers are transcendental, but the theory of p-adic L-functions shows that their "algebraic parts" are in fact congruent modulo p: we have $L(\psi_1^{12k+1}, 6k+1) = \dfrac{3^{9k-1}\,\Omega^{12k+1}}{(2\pi)^{6k}}\,\dfrac{C_k}{(6k)!}$ for some integer $C_k \in \mathbb{Z}$, and $S_p \equiv C_k \pmod p$. Now the calculation of C_k proceeds exactly like that of the number A_m defined in (108): C_k is, up to normalizing constants, equal to the value at $z = \mathfrak{z}_0$ of the $6k$-th non-holomorphic derivative of $\Theta(z)$, and can be computed quasi-recursively, and C_k is also equal to c_k^2 where c_k is, again up to normalizing constants, the value of the $3k$-th non-holomorphic derivative of $\eta(z)$ at $z=\frac{1}{2}(-1+\sqrt{-3})$ and is given by the formula in the theorem. Finally, an estimate of the size of $L(E_p,1)$ shows that $|S_p|<p$, so that S_p vanishes if and only if $c_k\equiv 0 \pmod{p}$, as claimed. The numbers c_k can also be described by a generating function rather than a quasi-recursion, using the relations between modular forms and differential equations discussed in §5, namely

$$(1-x)^{1/24}\,F\left(\frac{1}{3},\frac{1}{3};\frac{2}{3};x\right)^{1/2} = \sum_{n=0}^{\infty}\frac{c_n}{(3n)!}\left(\frac{x\,F(\frac{2}{3},\frac{2}{3};\frac{4}{3};x)^3}{8\,F(\frac{1}{3},\frac{1}{3};\frac{2}{3};x)^3}\right)^n,$$

where $F(a,b;c;x)$ denotes Gauss's hypergeometric function. ♡

References and Further Reading

There are several other elementary texts on the theory of modular forms which the reader can consult for a more detailed introduction to the field. Four fairly short introductions are the classical book by Gunning (*Lectures on Modular Forms*, Annals of Math. Studies **48**, Princeton, 1962), the last chapter of Serre's *Cours d'Arithmétique* (Presses Universitaires de France, 1970; English translation *A Course in Arithmetic*, Graduate Texts in Mathematics **7**, Springer 1973), Ogg's book *Modular Forms and Dirichlet Series* (Benjamin, 1969; especially for the material covered in §§3–4 of these notes), and my own chapter in the book *From Number Theory to Physics* (Springer 1992). The chapter on elliptic curves by Henri Cohen in the last-named book is also a highly recommended and compact introduction to a field which is intimately related to modular forms and which is touched on many times in these notes. Here one can also recommend N. Koblitz's *Introduction to Elliptic Curves and Modular forms* (Springer Graduate Texts **97**, 1984). An excellent book-length Introduction to Modular Forms is Serge Lang's book of that title (Springer Grundlehren **222**, 1976). Three books of a more classical nature are B. Schoeneberg's *Elliptic Modular Functions* (Springer Grundlehren **203**, 1974), R.A. Rankin's *Modular Forms and Modular Functions* (Cambridge, 1977) and (in German) *Elliptische Funktionen und Modulformen* by M. Koecher and A. Krieg (Springer, 1998).

The point of view in these notes leans towards the analytic, with as many results as possible (like the algebraicity of $j(\mathfrak{z})$ when $\mathfrak{z}$ is a CM point) being derived purely in terms of the theory of modular forms over the complex numbers, an approach which was sufficient – and usually simpler – for the type of applications which I had in mind. The books listed above also belong to this category. But for many other applications, including the deepest ones in Diophantine equations and arithmetic algebraic geometry, a more arithmetic and more advanced approach is required. Here the basic reference is

Shimura's classic *Introduction to the Arithmetic Theory of Automorphic Functions* (Princeton 1971), while two later books that can also be recommended are Miyake's *Modular Forms* (Springer 1989) and the very recent book *A First Course in Modular Forms* by Diamond and Shurman (Springer 2005).

We now give, section by section, some references (not intended to be in any sense complete) for various of the specific topics and examples treated in these notes.

1–2. The material here is all standard and can be found in the books listed above, except for the statement in the final section of §1.2 that the class numbers of negative discriminants are the Fourier coefficients of some kind of modular form of weight 3/2, which was proved in my paper in CRAS Paris (1975), 883–886.

3.1. Proposition 11 on the number of representations of integers as sums of four squares was proved by Jacobi in the *Fundamenta Nova Theoriae Ellipticorum*, 1829. We do not give references for the earlier theorems of Fermat and Lagrange. For more information about sums of squares, one can consult the book *Representations of Integers as Sums of Squares* (Springer, 1985) by E. Grosswald. The theory of Jacobi forms mentioned in connection with the two-variable theta functions $\theta_i(z,u)$ was developed in the book *The Theory of Jacobi Forms* (Birkhäuser, 1985) by M. Eichler and myself. Mersmann's theorem is proved in his Bonn Diplomarbeit, "Holomorphe η-Produkte und nichtverschwindende ganze Modulformen für $\varGamma_0(N)$" (Bonn, 1991), unfortunately never published in a journal. The theorem of Serre and Stark is given in their paper "Modular forms of weight 1/2" in *Modular Forms of One Variable VI* (Springer Lecture Notes **627**, 1977, editors J-P. Serre and myself; this is the sixth volume of the proceedings of two big international conferences on modular forms, held in Antwerp in 1972 and in Bonn in 1976, which contain a wealth of further material on the theory). The conjecture of Kac and Wakimoto appeared in their article in *Lie Theory and Geometry in Honor of Bertram Kostant* (Birkhäuser, 1994) and the solutions by Milne and myself in the Ramanujan Journal and Mathematical Research Letters, respectively, in 2000.

3.2. The detailed proof and references for the theorem of Hecke and Schoenberg can be found in Ogg's book cited above. Niemeier's classification of unimodular lattices of rank 24 is given in his paper "Definite quadratische Formen der Dimension 24 und Diskriminante 1" (J. Number Theory **5**, 1973). Siegel's mass formula was presented in his paper "Über die analytische Theorie der quadratischen Formen" in the Annals of Mathematics, 1935 (No. 20 of his *Gesammelte Abhandlungen*, Springer, 1966). A recent paper by M. King (Math. Comp., 2003) improves by a factor of more than 14 the lower bound on the number of inequivalent unimodular even lattices of dimension 32. The paper of Mallows-Odlyzko-Sloane on extremal theta series appeared in J. Algebra in 1975. The standard general reference for the theory of lattices, which contains an immense amount of further material, is the book by Conway and

Sloane (*Sphere Packings, Lattices and Groups*, Springer 1998). Milnor's example of 16-dimensional tori as non-isometric isospectral manifolds was given in 1964, and 2-dimensional examples (using modular groups!) were found by Vignéras in 1980. Examples of pairs of truly drum-like manifolds – i.e., domains in the flat plane – with different spectra were finally constructed by Gordon, Webb and Wolpert in 1992. For references and more on the history of this problem, we refer to the survey paper in the book *What's Happening in the Mathematical Sciences* (AMS, 1993).

4.1–4.3. For a general introduction to Hecke theory we refer the reader to the books of Ogg and Lang mentioned above or, of course, to the beautifully written papers of Hecke himself (if you can read German). Van der Blij's example is given in his paper "Binary quadratic forms of discriminant -23" in Indagationes Math., 1952. A good exposition of the connection between Galois representations and modular forms of weight one can be found in Serre's article in *Algebraic Number Fields: L-Functions and Galois Properties* (Academic Press 1977).

4.4. Book-length expositions of the Taniyama–Weil conjecture and its proof using modular forms are given in *Modular Forms and Fermat's Last Theorem* (G. Cornell, G. Stevens and J. Silverman, eds., Springer 1997) and, at a much more elementary level, *Invitation to the Mathematics of Fermat-Wiles* (Y. Hellegouarch, Academic Press 2001), which the reader can consult for more details concerning the history of the problem and its solution and for further references. An excellent survey of the content and status of Serre's conjecture can be found in the book *Lectures on Serre's Conjectures* by Ribet and Stein (http://modular.fas.harvard.edu/papers/serre/ribet-stein.ps). For the final proof of the conjecture and references to all earlier work, see "Modularity of 2-adic Barsotti-Tate representations" by M. Kisin (http://www.math.uchicago.edu/~kisin/preprints.html). Livné's example appeared in his paper "Cubic exponential sums and Galois representations" (Contemp. Math. **67**, 1987). For an exposition of the conjectural and known examples of higher-dimensional varieties with modular zeta functions, in particular of those coming from mirror symmetry, we refer to the recent paper "Modularity of Calabi-Yau varieties" by Hulek, Kloosterman and Schütt in *Global Aspects of Complex Geometry* (Springer, 2006) and to the monograph *Modular Calabi-Yau Threefolds* by C. Meyer (Fields Institute, 2005). However, the proof of Serre's conjectures means that the modularity is now known in many more cases than indicated in these surveys.

5.1. Proposition 15, as mentioned in the text, is due to Ramanujan (eq. (30) in "On certain Arithmetical Functions," Trans. Cambridge Phil. Soc., 1916). For a good discussion of the Chazy equation and its relation to the "Painlevé property" and to $\mathrm{SL}(2,\mathbb{C})$, see the article "Symmetry and the Chazy equation" by P. Clarkson and P. Olver (J. Diff. Eq. **124**, 1996). The result by Gallagher on means of periodic functions which we describe as our second application is described in his very nice paper “Arithmetic of means of squares and cubes” in Internat. Math. Res. Notices, 1994.

5.2. Rankin–Cohen brackets were defined in two stages: the general conditions needed for a polynomial in the derivatives of a modular form to itself be modular were described by R. Rankin in “The construction of automorphic forms from the derivatives of a modular form” (J. Indian Math. Soc., 1956), and then the specific bilinear operators $[\,\cdot\,,\,\cdot\,]_n$ satisfying these conditions were given by H. Cohen as a lemma in “Sums involving the values at negative integers of L functions of quadratic characters” (Math. Annalen, 1977). The Cohen–Kuznetsov series were defined in the latter paper and in the paper “A new class of identities for the Fourier coefficients of a modular form” (in Russian) by N.V. Kuznetsov in Acta Arith., 1975. The algebraic theory of “Rankin–Cohen algebras” was developed in my paper “Modular forms and differential operators” in Proc. Ind. Acad. Sciences, 1994, while the papers by Manin, P. Cohen and myself and by the Untenbergers discussed in the second application in this subsection appeared in the book *Algebraic Aspects of Integrable Systems: In Memory of Irene Dorfman* (Birkhäuser 1997) and in the J. Anal. Math., 1996, respectively.

5.3. The name and general definition of quasimodular forms were given in the paper “A generalized Jacobi theta function and quasimodular forms” by M. Kaneko and myself in the book *The Moduli Space of Curves* (Birkhäuser 1995), immediately following R. Dijkgraaf’s article “Mirror symmetry and elliptic curves” in which the problem of counting ramified coverings of the torus is presented and solved.

5.4. The relation between modular forms and linear differential equations was at the center of research on automorphic forms at the turn of the (previous) century and is treated in detail in the classical works of Fricke, Klein and Poincaré and in Weber’s *Lehrbuch der Algebra.* A discussion in a modern language can be found in §5 of P. Stiller’s paper in the Memoirs of the AMS **299**, 1984. Beukers’s modular proof of the Apéry identities implying the irrationality of $\zeta(2)$ and $\zeta(3)$ can be found in his article in Astérisque **147–148** (1987), which also contains references to Apéry’s original paper and other related work. My paper with Kleban on a connection between percolation theory and modular forms appeared in J. Statist. Phys. **113** (2003).

6.1. There are several references for the theory of complex multiplication. A nice book giving an introduction to the theory at an accessible level is *Primes of the form* x^2+ny^2 by David Cox (Wiley, 1989), while a more advanced account is given in the Springer Lecture Notes Volume 21 by Borel, Chowla, Herz, Iwasawa and Serre. Shanks’s approximation to π is given in a paper in J. Number Theory in 1982. Heegner’s original paper attacking the class number one problem by complex multiplication methods appeared in Math. Zeitschrift in 1952. The result quoted about congruent numbers was proved by Paul Monsky in “Mock Heegner points and congruent numbers”

(Math. Zeitschrift 1990) and the result about Sylvester's problem was announced by Noam Elkies, in "Heegner point computations" (Springer Lecture Notes in Computing Science, 1994). See also the recent preprint "Some Diophantine applications of Heegner points" by S. Dasgupta and J. Voight.

6.2. The formula for the norms of differences of singular moduli was proved in my joint paper "Singular moduli" (J. reine Angew. Math. **355**, 1985) with B. Gross, while our more general result concerning heights of Heegner points appeared in "Heegner points and derivatives of L-series" (Invent. Math. **85**, 1986). The formula describing traces of singular moduli is proved in my paper of the same name in the book *Motives, Polylogarithms and Hodge Theory* (International Press, 2002). Borcherds's result on product expansions of automorphic forms was published in a celebrated paper in Invent. Math. in 1995.

6.3. The Chowla–Selberg formula is discussed, among many other places, in the last chapter of Weil's book *Elliptic Functions According to Eisenstein and Kronecker* (Springer Ergebnisse **88**, 1976), which contains much other beautiful historical and mathematical material. Chudnovsky's result about the transcendence of $\Gamma(\frac{1}{4})$ is given in his paper for the 1978 (Helskinki) International Congress of Mathematicians, while Nesterenko's generalization giving the algebraic independence of π and e^{π} is proved in his paper "Modular functions and transcendence questions" (in Russian) in Mat. Sbornik, 1996. A good summary of this work, with further references, can be found in the "featured review" of the latter paper in the 1997 Mathematical Reviews. The algorithmic way of computing Taylor expansions of modular forms at CM points is described in the first of the two joint papers with Villegas cited below, in connection with the calculation of central values of L-series.

6.4. A discussion of formulas like (102) can be found in any of the general references for the theory of complex multiplication listed above. The applications of such formulas to questions of primality testing, factorization and cryptography is treated in a number of papers. See for instance "Efficient construction of cryptographically strong elliptic curves" by H. Baier and J. Buchmann (Springer Lecture Notes in Computer Science **1977**, 2001) and its bibliography. Serre's results on powers of the eta-function and on lacunarity of modular forms are contained in his papers "Sur la lacunarité des puissances de η" (Glasgow Math. J., 1985) and "Quelques applications du théorème de densité de Chebotarev" (Publ. IHES, 1981), respectively. The numerical experiments concerning the numbers defined in (108) were given in a note by B. Gross and myself in the memoires of the French mathematical society (1980), and the two papers with F. Villegas on central values of Hecke L-series and their applications to Sylvester's problems appeared in the proceedings of the third and fourth conferences of the Canadian Number Theory Association in 1993 and 1995, respectively.